NATIONAL OCCUPATIONAL EXPOSURE SURVEY

SAMPLING METHODOLOGY

Wm. Karl Sieber, Jr.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control
National Institute for Occupational Safety and Health
Division of Surveillance, Hazard Evaluations and Field Studies
Cincinnati, Ohio 45226

February 1990

DISCLAIMER

The contents of this report are expressed as received from the contractor.

Mention of a company name or product does not constitute endorsement by the National Institute for Occupational Safety and Health.

DHHS(NIOSH) Publication No. 89-102

FOREWORD

The National Occupational Exposure Survey (NOES) was a nationwide observational survey conducted in a sample of nearly 5,000 establishments from 1981-1983. The goal of the NOES was to compile data on the types of potential exposure agents found at the workplace, and the kinds of safety and health programs which had been implemented at the plant level. The sample of establishments included in the survey was designed to represent those segments of American industry covered under the Occupational Safety and Health Act of 1970.

This volume describes the method used to select the sample of plants to be surveyed, and the estimation techniques used to project survey data to national estimates.

I. ABSTRACT

The National Occupational Exposure Survey (NOES) of 1981-1983 was initiated by NIOSH to address a critical and continuing need for information on nationwide patterns of occupational exposures to potential health hazards. The NOES consisted of on-site observational surveys in a sample of nearly 5,000 establishments which had been selected to represent most sectors of the American workforce covered by the Occupational Safety and Health Act.

A two-stage sampling strategy was employed to construct the sample of establishments to be surveyed. The first stage resulted in the selection of 98 geographical areas, or primary sampling units. The geographical areas chosen in the first stage had relatively higher concentrations of those industries which were included in the target population. The second stage of sampling produced lists of establishments to be surveyed in the first-stage geographical areas. Establishments with 2,500 or more employees were not included in the first stage of sampling, and were treated separately in order to maintain more nearly equal probabilities of selection across establishments.

First stage selection of geographical areas was accomplished by random selection from strata defined by geography, number of employees, and concentration of establishments included in the target population. Second stage selection of establishments employed systematic sampling from a list of establishments ordered by number of employees and Standard Industrial Classification (SIC). The second stage sample was enlarged by 25 percent, and establishments in this enlarged sample were screened by telephone to determine eligibility for inclusion in the survey. A total of 4,490 establishments were ultimately surveyed in the NOES. Substitutions were made for establishments which fell outside the scope of the survey, and inspection warrants were obtained and enforced where necessary. The effective refusal rate among establishments selected for inclusion in the survey was 0.3 percent.

Two stages of ratio estimation were used in the process of projecting survey data to national statistics. Variances of the estimates were calculated using the method of balanced repeated replications.

CONTENTS

CONTENTS (Cont.)

FIGURES

TABLES

II. ACKNOWLEDGEMENTS

The National Occupational Exposure Survey was conducted by Westat, Incorporated, 1650 Research Boulevard, Rockville, Maryland, under NIOSH contract 210-80-0057.

I would like to thank the following people for their critical review of this document:

Mr. Wallace Carr
Project Officer, National Occupational Hazard Survey-Mining
Division of Respiratory Disease Studies
National Institute for Occupational Safety and Health

Mr. Robert Hanson
Mr. John Edmonds
Ms. Diane Ward
Westat, Incorporated

Dr. Kathryn R. Mahaffey
Chief, Priorities and Research Analysis Branch
Division of Standards Development and Technology Transfer
National Institute for Occupational Safety and Health

Mr. Stanley Shulman
Mathematical Statistician
Division of Physical Sciences and Engineering
National Institute for Occupational Safety and Health

Mr. Gary Shapiro
Assistant Division Chief
Income, Longitudinal and Expenditure Survey Design
Statistical Methods Division
Bureau of the Census

I am grateful to Mr. Todd Frazier, Mr. David Sundin, and Mr. David Pedersen who each provided extensive comments as to the organization and content of this report. Mr. Randy Young provided information on computer operations during the National Occupational Exposure Survey. Mrs. Kathy Mitchell showed special patience in preparing the several drafts of this report.

III. INTRODUCTION

The National Institute for Occupational Safety and Health (NIOSH) is charged with developing information on the types and extent of exposures to occupational health hazards (1). To develop data of this type, NIOSH has carried out two on-site observational surveys of a sample of facilities representative of selected segments of American workplaces. The first, the National Occupational Hazard Survey (NOHS), was conducted by NIOSH from 1972 to 1974. The second was the National Occupational Exposure Survey (NOES) conducted by NIOSH between 1981 and 1983. To a great extent, the NOES was designed to provide results which could be compared to those obtained from the NOHS.

The NOES is a response by NIOSH to the continuing need for information on nationwide patterns of occupational exposure to health hazards. This report, second in a series of reports based on the NOES, details the development of the sample design, selection of sample establishments, and the statistical methodology developed to make national projections from data obtained by surveying a probability sample of worksites and potential workplace hazards. Volume I, National Occupational Exposure Survey - Survey Manual, detailed field guidelines and the actual questionnaire used in the NOES (2).

In summary, the objectives of the NOES were:

1. For selected industrial sectors, to develop estimates of the number of workers potentially exposed to chemical, physical, and biological agents;

2. To develop data that describe the nature and extent of these potential exposures to health hazards and the degree to which businesses have implemented programs to reduce occupational health problems; and

3. To compile data such that analysis of industrial hazard exposure trends would be possible by comparison with similar data collected in NOHS.

The target population was defined as employees working in establishments or job sites located in the United States reporting eight or more employees at the time of the survey, and with a primary activity or line of business on a list of target Standard Industrial Classification codes (SICs) (3). An establishment was defined as an economic unit, generally at a single location, where business, service, or industrial activities were performed.[1]

Development and implementation of the NOES sample design was done under NIOSH contract no. 210-80-0057 to Westat, Incorporated, Rockville, Maryland.

---

[1] The terms establishment, facility, firm, and worksite are used interchangeably in this report.

A. Development of the 1972-1974 Survey: The NOHS

Sampling methodology in the NOES was generally based on a design used for the NOHS. The NOHS involved a two-stage selection procedure with stratification. In that survey, primary sampling units (PSUs) were defined from the 247 Standard Metropolitan Statistical Areas (SMSAs) in 1970 and certain urban areas. Each PSU defined a geographic cluster of business and industrial establishments. The sample consisted of 67 PSUs selected with probability proportional to size, i.e., proportional to the number of establishments in defined strata.

Establishments within the 67 PSUs were stratified by probability of selection of the PSU, number of employees, and SIC code, and individual establishments were selected for field interview using systematic selection. A sample size of 4,636 facilities was determined in this manner.

Further details concerning sample selection and analysis for the NOHS is described in Volume II of the NOHS series, 'Data Editing and Data Base Development' (4).

B. 1981-1983 NOES Sampling Strategy

The NOES also used a two-stage sampling strategy for most of the sample. The first stage of sampling involved selection of a sample of 98 PSUs. PSUs in the NOES were defined across all 50 states rather than only as SMSAs as was done in the NOHS. This is one reason why more PSUs were selected in the NOES than in the NOHS. With the exception of samples of very large establishments drawn irrespective of geographic location, the interviewed sample was confined to these 98 PSUs. Stratification of PSUs was based on geography, number of employees, and concentration of establishments operating in select industries. The second stage (within PSU) selection for establishments was done using a systematic selection procedure. Very large establishments (2,500 or more employees) were treated separately in order to maintain more nearly equal probabilities of selection across establishments.

The SIC codes of firms eligible for this survey are shown in Appendix A. Establishments with eight or more employees and conducting business within this specific set of SICs (called the "target SICs") were considered to be in-scope in the NOES. Establishments with eight or more employees only were considered for comparability with the NOHS and because accurately surveying establishments with less than eight employees would have been difficult. Coverage of construction and manufacturing establishments was emphasized in the NOES by defining these SIC categories to include a broad range of SICs, while finance establishments as well as mining and mineral processing establishments were excluded from it.

The interviewed sample was designated in two steps: (1) a sample of 7,392 establishments was contacted by telephone to identify those establishments that were in the scope of the study; and (2) those establishments identified in (1) were visited and surveyed. A total of 4,490 establishments had complete field interviews in the NOES.

Figure 1 is an outline of the sampling strategy followed in the NOES. A total of 604 PSUs were defined for the sampling process. PSUs were defined geographically with the county as the primary unit. Some PSUs consisted of a single county, e.g., Orange County, California. Other PSUs were made up of counties that constituted a SMSA in 1980, which in a few cases crossed state boundaries: e.g., Cincinnati SMSA consisting of Dearborn County, Indiana; Boone, Campbell, and Kenton Counties in Kentucky; and Brown, Clermont, Hamilton, and Warren counties in Ohio. The 604 PSUs included 446,125 establishments eligible for the survey.

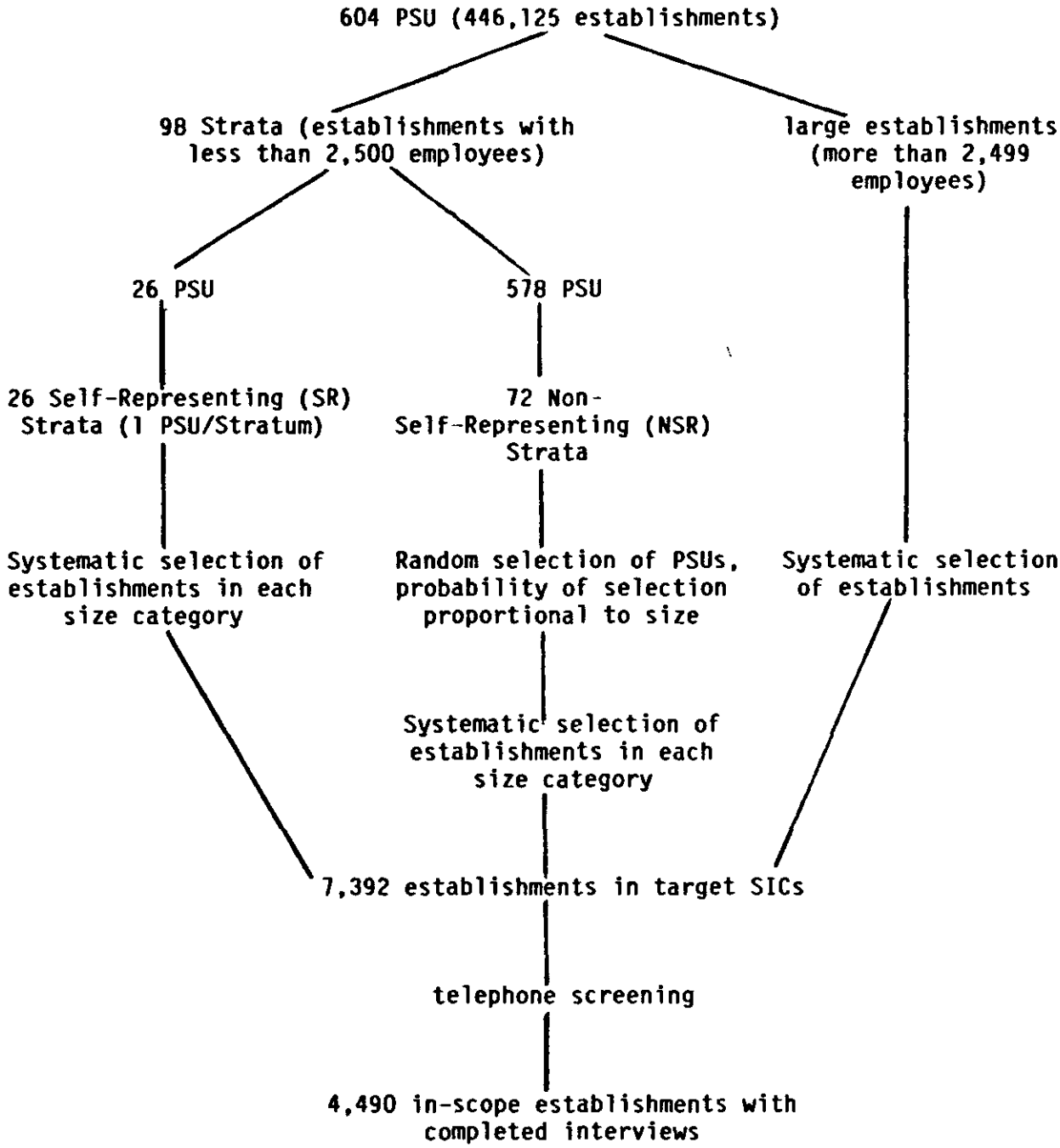
The 604 PSUs were stratified into 98 strata. The purpose of stratification was to obtain groups of PSUs which were of equal size and were homogeneous with respect to variables of interest in the NOES. Some of the criteria used for designing strata included: proportion of employees in firms where high potential exposure to health hazards might be found (e.g., chemical, rubber, or leather industries), geography (census region), and SMSA or non-SMSA. The 98 strata consisted of 26 self-representing (SR) strata, made up of 1 large PSU each, and 72 non-self-representing (NSR) strata made up of the remaining 578 PSUs.

The selection of establishments with less than 2,500 employees was done from 98 PSUs, one from each of the 26 SR and 72 NSR strata. These 98 PSUs are listed in Appendix B. A systematic sample of establishments in each of the 26 PSUs making up the SR strata was designated to be interviewed. Samples were selected independently across establishment size classes, where size was defined as the size of the workforce at that work site. PSUs in the 72 NSR strata from which establishments were to be selected were chosen as a random sample with probability of selection proportional to the number of establishments contributed by that PSU to that stratum, i.e., the measure of size of that PSU. One PSU was chosen from each NSR stratum. Systematic selection of establishments in each NSR PSU was then done using methods identical to those for selecting establishments from SR PSUs.

The sample of establishments employing 2,500 or more employees was designated without regard to sampling from PSUs. Samples in each of the size categories with these employee levels were determined using systematic selection of all firms nationwide with 2,500 or more employees.

Sample establishments were contacted by telephone to confirm that those establishments had enough employees and operated in an appropriate SIC to be included in the survey, and would participate in it. This sample was known as the 'screening' sample and consisted of 7,392 establishments. After screening, 4,504

FIGURE 1. OUTLINE OF SAMPLING STRATEGY
NOES 1981-1983



establishments were designated for field interview of which all but 125 of these establishments were interviewed. Substitutes were found for 111 of the 125, making the total number of completed interviews 4,490. The effective refusal rate of establishments for participation in the NOES was .3 percent.

Two stages of ratio estimation were used in the estimation process. Variances of estimates were calculated using the method of balanced repeated replications.

Much of the sample selection was carried out as a computer operation. National estimates were also determined using a computer software package.

IV. SAMPLE DESIGN

Listings from the Bureau of the Census publication County Business Patterns - 1978 (CBP) provided data needed to establish sampling rates, while listings from the 1980 Dun and Bradstreet Market Inventory (DMI) were used to select establishments. Supplementing these lists for completeness was considered, but was not done because of the costs involved and extensive coverage of the DMI. The initial screening operation was done to select a sample of establishments employing eight or more employees and operating in one of the Standard Industrial Classification (SIC) (3) codes covered by the NOES. This screening was carried out as a telephone survey which identified establishments still in business and eligible for the survey during the 1981-1983 data collection period. The sampling plan attempted to produce minimum variance for a fixed cost by considering strata determined by number of employees at the worksite.

The design of NOES made use of prior experience from the NOHS. The NOHS data provided guidance as to the method of stratification and most efficient sampling rates in strata.

A. Sources of Data for the Sampling Frame

The design of NOES was based on information from the Bureau of Census publication County Business Patterns, 1978 (CBP) (5). The CBP was used to estimate the number of establishments and size of the workforce in establishments in each PSU. Information on individual establishments' size and location was supplied by the Dun and Bradstreet Market Inventory (DMI) (6). The DMI is a well-known and widely used industrial directory service. Historically businesses were listed in the file so as to establish credit ratings. Thus the list does not represent all U.S. industries. A special effort has been made by Dun and Bradstreet to expand the DMI file in order to have more complete listings of establishments, however, and the DMI is considered close to complete.

An examination of the completeness of the DMI was made before deciding on its use in the NOES. Establishments in the following SIC groups were found to have DMI to CBP employee ratios of less than 0.9; i.e., presumably ten percent under-representation was found in the DMI file:

- 451 & 452 - Air transportation
- 481 - Telephone communication
- 491 - Electronic services
- 493 - Combination electric, gas and other services combined
- 5541 - Gasoline service stations
- 7231 - Beauty shops
- 7241 - Barber shops
- 7299 - Miscellaneous personal services

Since supplementing the DMI list to cover these SICs was considered beyond available resources and the DMI already was quite extensive in coverage of most SIC groups included in the survey, the coverage provided by the DMI was considered without supplementation. See Appendix C for more discussion on this point.

B. Defining the Target Population

The target population was defined as those establishments or job sites located in the 50 states reporting eight or more employees and having as a primary activity one of the target SICs listed in Appendix A.

As is the case with any sample survey, inferences from the sample data are restricted to the target population. The following points provide a description of the target population.

Establishments included:

- Those located in metropolitan and other urbanized areas of the United States in 1980 and which were still worksites during the 1981 to 1983 period of data collection.

- Those reporting eight or more employees in the 1978 CBP and 1980 DMI files, provided that these establishments were still in business and operating during the period of data collection.

Establishments excluded:

- Establishments engaged in agricultural production, any mining activity except oil and gas extraction, railroad transportation, private households, finance institutions, and all Federal, State, and municipal government facilities.

Within each PSU, establishments were classified by number of employees. Eleven size classes were defined as follows: 8-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1000-1499, 1500-2499, 2500-4999, and 5000+ employees, and those for which employment totals were not available from the DMI. The two largest categories were treated separately from the others, since they represented a substantial expenditure of time for the surveyor teams and would affect calculation of variances of the survey results.

C. Derivation of the Sample Design

Methods of optimizing the sample design for a survey typically involve establishing a cost function for the study, expressing the sampling variance, and solving the equation which will produce the minimum variance for a fixed cost (7). This approach was an oversimplification of the needs for the NOES because it assumed there was a single statistic whose variance is to be minimized. There were several different types of statistics for which estimates were needed

from this survey and quite different sample designs could have been chosen depending on which statistic was considered to be of greatest importance.

Much of the analysis in the 1972-1974 NOHS referred to industry-by-industry breakdowns. For these kinds of analyses, the samples of industries should have approximately equal reliability. This would lead to a sample design with roughly equal sample sizes by industry. On the other hand, an efficient sample for analysis of statistics for all industries combined would require that the sample size in each industrial sector be proportional to that sector's contribution to the total number of establishments eligible for the survey.

A second problem arises from the interest in data on the distributions of both establishments and employees. An efficient sample design for statistics on employees would use higher sampling rates for larger establishments than for smaller ones. For statistics on establishments, however, the number of plants, rather than their size, would be important.

The sample design developed for NOES maximized the reliability of estimates of numbers of employees. Although estimates of facilities are available using the methodology developed in NOES, breakdown by industry or data on the number of firms with specific characteristics was assigned lower priority in developing the sample design.

1. The Cost Function

   A cost function expressing the total cost as the sum of costs over employee size strata was first determined. The cost within a size stratum was equal to the product of the number of sample establishments and the average cost of interviewing the establishments. Average costs were expressed as number of person-hours required for that size group.

   This cost function recognizes only the unit costs and the total cost for those aspects of the survey that are directly affected by the sample size. The number of PSUs does not enter the cost function. There are several reasons for this. First, the cost of designating the sample of establishments, a major portion of which would involve the use of the telephone, would be directly related to the number of sample establishments and would have little relationship to the number or location of the PSUs. Second, the cost of surveyor travel between PSUs was assumed to be relatively small and concentrated during weekends and would not influence the total cost based on person hours during the regular week. Third, the time a team would be assigned to a PSU was restricted to below a 30-day maximum because of Federal government restrictions on per diem reimbursement. These conditions meant the number of PSUs would have little impact on the total survey cost, and so the number of sample PSUs was set as large as administratively feasible. A large number of PSUs also reduces the component of variance arising from the sampling of PSUs.

The cost function was defined as:

$$C = \sum_a n_a C_a \qquad (1)$$

where

C = The total cost.

C_a = The cost per sample establishment in the a^{th} employer size class.

a = Employee size class, 1 to 10, see Table 1.

n_a = The number of sample establishments selected in the a^{th} size class.

The term C_a in the cost function is the total number of person hours of surveyor time per establishment in the survey in the a^{th} size class. These unit costs varied according to the size of the establishment and were taken from a tabulation of average surveyor hours per establishment by size class experienced in the NOHS. They are listed by employee size class in Table 1. It was assumed the amount of time required to survey a sample firm would be similar to the experience in the NOHS. The total of all costs of the survey also included a number of more-or-less fixed charges that did not vary directly with moderate changes in the sample size; for example, writing specifications and computer programs for data processing, overhead costs, the cost of hiring and training surveyors, etc.

The total of all directly related costs of the sample was taken as the total number of paid person hours to support a proposed number of surveyors working for an expected survey period of two years. This ignored the cost of travel and all fixed costs. It also meant that other costs expected to vary with the sample size were assumed to be small and not important in determining the sample size. For example, the total cost of telephone screening was assumed not to be importantly affected by variations in the allocations of the sample among size classes, or by minor changes in the number of sample cases.

An average number of 21 surveyors were expected to be available for the survey period. The surveyors were to be assigned to five teams, and each team was to have a team leader. Because of time spent in team supervision, each leader was assumed to produce 80 percent of the production of the other team members. Each team member was to work a 40 hour week for 48 weeks of the year; the remaining four weeks were to be taken up by annual and

TABLE 1. SAMPLING RATES AND EXPECTED DISTRIBUTION OF SAMPLE OF ESTABLISHMENTS
NOES 1981-1983

Employee size class	Number of employees	Number of establishments[1] N_a	Survey cost (hrs.)[2] C_a	Average number of employees per facility	Relvariance factors	Sampling interval k_a	oversampling ratio f_a	Facilitie in sample n_a
1	8-19	237,445	3.18	12.4	2	199.530	1.0	1,190
2	20-49	114,508	4.28	32.4	1	125.250	1.593	914
3	50-99	44,567	5.82	71.7	1	66.030	3.022	675
4	100-249	30,601	8.66	158.1	1	36.520	5.464	838
5	250-499	10,887	14.36	349.9	1	21.260	9.384	512
6	500-999	5,055	26.79	690.1	1	14.700	13.576	344
7	1000-1499	1,424	49.78	1,200	1	11.580	17.235	123
8	1500-2499	906	66.52	1,900	1	8.389	23.785	108
9	2500-4999	520	86.16	3,500	1	5.545	35.984	94
10	5000+	212	189.25	9,250	2	2.190	91.110	97
11	Not Available	Unknown	Unknown	Unknown	1	199.530	1.000	Unknown
Total		446,125						4,895

[1] Based on tabulation of CBP county summary records for 1978.

[2] Person hours per establishment required to investigate facilities in NOHS.

sick leave and holiday time. For a 24 month data collection period, the hours contributed by the five team leaders would be:

(5 leaders) x (40 hours) x (48 weeks) x (2 years) x (.8) = 15,360 hours,

and from the remaining 16 team members:

(16 members) x (40) x (48) x (2) = 61,440 hours.

The total for all 21 surveyors = 76,800 person hours.

These assumptions turned out to be only a rough approximation of the actual survey conditions. The period of data collection was about 32 months rather than the predicted 24. Also, the survey force began initially with only 11 surveyors, rose to 15 after 5 months, and then fluctuated between 10 and 22 for most of the remaining survey period. The size of the field staff meant that surveyor teams did not function as expected. Costs in terms of hours to survey plants for NOES were also found to differ from the NOHS experience. The actual costs per establishment size class are detailed in Volume I of this series (2).

2. The Variance Function

The variance for the estimated total number of establishments was taken as:

$$\sigma^2(Y') = \sum_a N_a^2 \, (1/n_a - 1/N_a) \, S^2(\bar{Y}_a) \qquad (2)$$

where

Y' = a total, estimated from the survey.

N_a = The total number of establishments in the universe of study in the a^{th} establishment size class.

n_a = The number of establishments in the sample from the a^{th} size class.

$S^2(\bar{Y}_a)$ = The estimated population variance of the number of establishments with the characteristic y in the a^{th} size class.

The estimated variance $S^2(\bar{Y}_a)$ for the a^{th} size class is given by:

$$S^2(\bar{Y}_a) = \sum_{i=1}^{N_a} \frac{(Y_{ai} - \bar{Y}_a)^2}{(N_a - 1)} \qquad (3)$$

where

Y_{ai} = The number of employees having the characteristic Y in the i^{th} establishment in the a^{th} size class.

$\bar{Y}_a = (\sum_{i=1}^{N_a} Y_{ai})/N_a$ is the average number of employees with the characteristic per establishment in the a^{th} class.

Values of $S^2(\bar{Y}_a)$ were not available when the sample design was developed. Although data from NOHS could have been used to estimate the values of $S^2(\bar{Y}_a)$ for a selected set of characteristics, the time schedule prevented waiting for these variances to be prepared. Instead, an approximation in which the relvariances of desired characteristics were assumed to be constant within most size classes was employed (8, 9). This approximation was based on experience in other surveys.

With this assumption:

$$\text{relvariance} = \frac{S^2(\bar{Y}_a)}{(\bar{Y}_a)^2} = \text{constant}$$

$$S^2(\bar{Y}_a) = \text{constant} \times (\bar{Y}_a)^2 \qquad (4)$$

If the value of the constant and mean number of employees in size class a with characteristic Y are known, an approximation to the variance $S^2(\bar{Y}_a)$ for the a^{th} size class can be made. The assumption of a constant relvariance is weakest in the largest and in the smallest size classes, and so the constant was doubled for these classes. Values of the constant are also shown in Table 1.

The variance expression does not include the contribution to the variance that arises because most of the sample was restricted to the 98 sample PSUs. The between PSU variance did not need to appear in calculations for optimum sample size since, because the cost function did not account for total number of PSUs, it had been decided to have as many sample PSUs as possible, and so minimize that component of variance resulting between PSUs.

Optimum allocation of facilities selected in the a^{th} size class is that sample size which would produce minimum variance at the fixed cost. The equations involved and methods of solution are outlined in Appendix D. The optimum sample size to be selected from the a^{th} size class is given by:

$$n_a = \frac{N_a S(\bar{Y}_a)}{[\sum_a N_a S(\bar{Y}_a)\sqrt{c_a}]} \times \frac{C}{\sqrt{c_a}} \qquad (5)$$

where all quantities are as defined above.

If values of $S(\bar{Y}_a)$ from expression (4) are substituted in equation (5), the variance constant terms cancel out. Optimum sample sizes for size class a then involve only the relative sizes of the variance constants for size class a, the mean number of employees with characteristic Y, the number of establishments, and unit costs.

Since the CBP provided the most precise estimates of the current number of establishments and employees in the target SICs, it was used to determine N_a and $\bar{Y}_a$. In some cases adjustment of $\bar{Y}_a$ was necessary, however. The values of C_a were estimated from NOHS. The size classes used in the CBP records did not permit the size classifications defined earlier, so approximations of the CBP counts for the correct size classes were obtained by using the proportions of the establishments that appeared in those size classes in the DMI file.

Since N_a was based on total numbers of CBP establishments, the values of n_a that define optimum sample size were given in terms of CBP establishments. However, the important result of the optimization computations was to find the optimum sampling rates n_a/N_a for establishments in the a^{th} size class. These rates could then be applied to the DMI file. Parameters used in selecting the sample, and the expected number of CBP establishment selections resulting from the optimum sampling rates are given in Table 1. Numbers of establishments in each size class expected from both the CBP and DMI files are shown in Table 3 in Chapter VI.

The actual samples from the DMI were expected to differ somewhat from the expected sample totals derived from the CBP universe (see Table 3). A number of other factors also affected sample sizes in the DMI. When the computed sampling rates were applied to DMI universe files having duplicate records for establishments or having records for establishments no longer in business, the usual result was a larger sample than expected; however, the telephone screening operation eliminated the out-of-business sample cases. Similarly, when the incomplete file was sampled with these rates, inadequate coverage was reflected by a corresponding shortage in the number of cases selected. The DMI file was expected to have both under-coverage and multiple listing problems.

V. PRIMARY SAMPLING UNITS (PSUs)

Geographical and surveyor workload restrictions resulted in defining 604 Primary Sampling Units (PSUs). PSUs were made up of contiguous counties, parishes in Louisiana, census divisions in Alaska, and in metropolitan areas were composed of counties that made up Standard Metropolitan Statistical Areas (SMSAs). The 604 PSUs were stratified into 98 strata. Of these strata, 26 were grouped into self-representing strata with one large PSU per stratum. The remaining 578 PSUs were grouped into 72 strata, called non-self-representing strata. One PSU from each non-self-representing stratum was selected to represent all other PSUs in the stratum. This selection was done with probability proportional to size. A total of 98 PSUs were selected for analysis in the NOES. Sample establishments with less than 2,500 employees were selected from these 98 PSUs, while establishments with 2,500 or more employees were selected using systematic selection across all 604 PSUs.

A. Definition of Primary Sampling Units

The county was the basic building block for PSUs. This was done to enable the telephone interviewer and the surveyor to use a familiar boundary for a PSU. The DMI file used for sampling establishments also records the county location for establishments as part of the address information.

The system for defining individual PSUs was also heavily influenced by the expected organization of the surveyor staff and the number of surveyors expected to be available for conducting interviews at sample establishments. Originally, a staff of 21 trained surveyors working in five teams was expected to conduct field interviews over a period of two years.

All counties in the 50 States and the District of Columbia were combined into 604 PSUs for this survey. Several conditions were important in defining the PSUs:

1. PSUs as Combinations of Counties

   PSUs were made up of contiguous counties. In Louisiana and Alaska, parishes and census divisions, respectively, took the place of counties. Independent cities were combined with neighboring counties.

2. Metropolitan PSUs

   PSUs in metropolitan areas were made up of the counties that composed SMSAs at the time of the 1980 census. In some smaller SMSAs, additional non-metropolitan counties were added to provide sufficient interviewing workloads.

3. Non-metropolitan (Non-metro) PSUs

   PSUs in non-metro areas were made up of groups of counties that had common boundaries. These counties or groups of contiguous

counties were large enough so that a self-weighting sample of the size planned and which would provide a sufficient surveyor workload could be selected.

4. State Boundaries and PSUs

   Although the intent was to construct PSUs from counties within the same state, multi-state PSUs did occur either because some SMSAs included areas in more than one state, or because the algorithm used to assign non-metro counties to PSUs occasionally included parts from more than one state.

5. Surveyor Workloads

   Each PSU was constructed to provide enough sample establishments to keep a four-person surveyor team busy for a period of two to four weeks.

B. Establishing the Size of the PSUs

The sample was designed to incorporate a self-weighting sample within employee size classes. A self-weighting sample was determined by considering the probability of selecting a specific establishment. The overall probability of selecting an establishment is equal to (the probability of selecting the PSU containing the establishment) times (the probability of selecting the establishment from the selected PSU). For the size class having the lowest sampling rate (i.e., the size class with the greatest numbers of establishments), the self-weighting sample was defined by the following condition:

$$\frac{1}{k} = \left(\frac{M_{hj}}{M_h}\right) \times \left(\frac{M_h}{M_{hj}} \times \frac{1}{k}\right) \qquad (6)$$

where

M_{hj} = The total number of establishments for the survey in the j^{th} PSU and h^{th} stratum, i.e., $\sum_a N_{hja} f_a$, the measure of size of the j^{th} PSU in the h^{th} stratum.

M_h = The total number of establishments in the h^{th} stratum, i.e., $\Sigma_j M_{hj}$, the measure of size of all PSUs in the h^{th} stratum.

N_{hja} = The number of establishments in the U.S. in the a^{th} employee size class (according to CBP) in the j^{th} PSU in the h^{th} stratum.

f_a = The oversampling ratio for establishments in the a^{th} size class (see below).

k = Sampling interval, $1/(n_a/N_a)$.

This expression is derived in Appendix E. The term (M_{hj}/M_h) on the right of expression (6) is the probability of selecting the sample PSU from among all PSUs in its stratum. The remaining term on the right defines the probability of selecting sample establishments from the sample PSU. For the a^{th} size class, the following general expression defines the sampling system:

$$\frac{f_a}{k} = \left(\frac{M_{hj}}{M_h}\right) \times \left(\frac{M_h}{M_{hj}} \times \frac{f_a}{k}\right) \tag{7}$$

where f_a is the oversampling ratio for establishments in the a^{th} size class. The oversampling ratio is the ratio of the largest sampling fraction to the sampling fraction in the a^{th} size class (see Chapter VI).

The terms on the right of expression (7) have the same meaning as in (6); the probability of selection of the PSU is the same but the probability of selection of establishments within the PSU reflects the larger overall sampling fraction f_a/k that applies to the a^{th} class.

These conditions gave rise to two restrictions which can be expressed statistically. The within PSU selection probability given in (7) is the basis of one condition:

1. The probability of selection of establishments within PSUs should not exceed 1; that is:

$$\left(\frac{M_h}{M_{hj}}\right) \times \left(\frac{f_a}{k}\right) \leq 1$$

It follows that the PSU measure of size must satisfy:

$$M_{hj} \geq \frac{M_h f_a}{k} \tag{8}$$

This restriction was imposed so that a self-weighting sample could be obtained. Writing $\tilde{f}_{ahj}$ as the value of f_a for the largest class with an establishment in the hj^{th} PSU enabled a lower bound to be placed on the measure of size for the hj^{th} PSU:

$$M_{hj} \geq \frac{M_h \tilde{f}_{ahj}}{k} \tag{9}$$

2. At least two team weeks of effort should be required to survey the sample expected from the PSU. Expressed algebraically, this condition becomes:

$$2(139.7) \leq \left(\frac{M_h}{M_{hj}}\right) \sum_{a=1}^{8} \left(\frac{f_a}{k}\right) x\ (N_{ahj}C_a) \qquad (10)$$

where C_a is the per firm surveyor hours for the a^{th} class and 2(139.7) hours of productive surveying per two week period were expected from each four person surveying team. The right side of this expression shows the total surveyor hours in the hj^{th} PSU as the sum of the products of the number of sample firms in the classes and the per firm survey hours needed. The number of hours of productive surveying per week was derived as follows:

		Hours Per Week
Supervision (40 hours x .2)	=	8.0
Leave (4 persons x 40 hours x 4/52 fraction of weeks in leave status)	=	12.3
Investigation (remaining hours of week)	=	139.7
Total (4 persons x 40 hours)	=	160.0

Condition 2 was used to define an upper limit on the PSU measure of size as:

$$M_{hj} \leq \left(\frac{M_h}{2(139.7)k}\right) \sum_{a=1}^{8} f_a N_{ahj} C_a \qquad (11)$$

Although conditions 1 and 2 could be stated explicitly, it was not always practical to adhere to them rigidly. For example, the PSU measure of size M_{hj} for some PSUs could be made large enough to satisfy condition 1 only by defining PSUs covering excessively large areas, and, some PSUs had to be defined with measures that did not meet this condition. This problem occurred for employee size classes 3 through 8 in these PSUs and was dealt with by assigning special weights in the estimation procedure (see Chapter VIII).

C. Location and Stratification of PSUs

The grouping of U.S. counties into 604 PSUs for NOES was done in a series of manual and computer assisted steps following the conditions specified in Section A and conditions 1 and 2 of Section B.

All counties, parishes, and independent cities within the United States were listed in a contiguous sequence. The list was prepared by manually assigning sequence numbers to the counties on a series of maps. The ordering tried to minimize "cross-overs" from one side

of a significant geographical feature to the other and from a state to its neighbors. Particular care was taken to minimize cross-overs from one Census Region to another (i.e., Northeast, North Central, South and West, as designated by the hundreds digit of the PSU number, 2, 4, 6, 8 respectively as shown in Appendix B).

SMSAs were defined as PSUs. In a few instances, one or more adjacent non-metro counties were added to smaller SMSAs to obtain minimum PSU sizes. This was done in a way to minimize 'cross-overs'

Although each of the very large SMSAs was treated as a single PSU, field interviewing was occasionally apportioned to more than one team to be done at different times. For example, one-half of the Chicago SMSA was surveyed by all of the interviewers available at the start of the survey and the remaining portion of the Chicago SMSA was interviewed later as a separate assignment.

Non-metro counties were combined into PSUs following the two conditions for size discussed in Section B. This step was done using a computer. The computer results were visually inspected to look for awkward geographical combinations that would make them inappropriate assignment areas. A few PSUs of very large area were generated in the Western states. Counties in these states were resequenced and a revised set of PSUs were generated. Later, when one of these large PSUs (in Alaska) was identified as a sample PSU, a subsample of the PSU area was selected to permit manageable travel patterns.

Stratification of the PSUs was imposed so that data from the many exposure groups included in the Survey could be handled easily. PSUs with similar characteristics such as number of employees or proportions of employees in certain industries were grouped and treated as a unit in the process of stratification. The 604 PSUs defined in the NOES were grouped into 98 strata. Selection of establishments was then done from 98 PSUs within the 98 strata, rather than from all 604 PSUs. Stratification also reduces the variance between PSUs within each stratum. The efficiency of a stratified design as measured by the variance is improved by defining strata of approximately equal size such that the PSUs within the strata are as homogeneous as possible with respect to the important statistics to be estimated from the survey. Homogeneity of PSUs within strata can sometimes be improved by using groups of PSUs with similar economic structure as strata.

The requirement that PSUs should provide an interviewing assignment of two to four weeks for a four person surveyor team was an important consideration in determining the size of the strata. The number of sample establishments and the average survey cost per establishment shown in Table 1 in Chapter IV indicate that the expected number of surveyor hours for establishments with 2,500 or more employees should have been about 35 percent of the total surveyor workload. As these large establishments were to be surveyed without regard to their location, they did not influence the number of sample PSUs. Strata sizes were therefore based on

apportioning the remaining 65 percent of the 75,200 hours (about 48,400 person hours) required to survey establishments with less than 2,500 employees.

Given 139.7 hours of productive surveying per team, per week, the total number of team weeks for surveying establishments of less than 2,500 employees approximates:

48,400/139.7 = 345 team weeks

Assuming two to four weeks of surveying time per PSU, the average workload over all PSUs should be three team weeks. Then an approximate duration (in terms of hours spent surveying) for each stratum would be 3 weeks out of 345 (or about 1 in 115) of the total survey workload for establishments of less than 2,500 employees.

The disparity in size of the PSUs interfered with establishing strata of equal sizes because some PSUs were larger than the desired average stratum size. The largest of these PSUs were defined as separate strata (self-representing strata) and the remaining PSUs were grouped into strata of approximately equal size.

Sample establishments with less than 2,500 employees were to be designated from PSUs within each strata. Since very large establishments with 2,500 or more employees were to be selected without regard to their PSU location, that did not influence the stratification process.

PSUs in the strata should also be relatively homogeneous with respect to statistics of interest for the survey. Groups of PSUs with significant concentrations of employees in certain key target industries that were likely to have serious and common health hazards were identified. This worked fairly well for most of the small PSUs. However, for PSUs which contained a wide range of target industries, it was not always possible to produce strata that were homogeneous in this regard. This was particularly true in the larger employee size classes. Additional stratification criteria, in addition to employee concentration by SIC, were therefore used. The computer was used to group PSUs and display the distribution of PSUs with respect to the following variables:

1. Proportion of employees in establishments working in manufacturing SICs.

2. Proportion of employees in establishments within the PSU falling in the largest size classes.

3. Concentration of employees in the petroleum and/or chemical, rubber, leather industries.

4. Geography - Census region.

5. SMSA or Non-SMSA.

This listing also reflects the order of importance of each variable in the formation of strata. As a first step, large groups were formed comprising PSUs that were similar with respect to numbers of employees in manufacturing and large establishments. If possible, employees in industries thought to have high potential exposures, e.g., petroleum, chemical, rubber, and leather industries, were also similarly concentrated. The measure of size in each large group determined the number of strata that should be produced from the group. If two or more strata were to be constructed, the PSUs were sorted by the five variables in the order listed above and then divided into strata, based upon total measure of size and the similarity across PSUs for each variable above.

The process produced a total of 98 strata. Twenty-six of these strata contained only one large PSU; these strata are called self-representing (SR) because the single PSU represents itself in the sample. The remaining 578 PSUs (604 minus 26) were grouped into 72 non-self-representing (NSR) strata having about equal measures of size; the term NSR was applied to these strata because one PSU was selected to represent all other PSUs in its stratum. The final groupings of PSUs into strata were done by the contractor, and are not available.

D. Selection of Sample PSUs

Once the strata were defined, all PSUs were listed by stratum showing M_{hj}, the PSU measure of size. Prior to sampling, the strata were compared to locate pairs of strata that were composed of roughly similar PSUs. The pairing of strata was significant, since the computation of variances described in Chapter VIII employed a paired stratum method.

One PSU was selected at random from each stratum with the probability of selection for each PSU proportional to the measure of size contributed by that PSU. The composition of the 98 PSUs selected for the NOES is shown in Appendix B. Parts of 40 States and the District of Columbia appear among the sample PSUs.

VI. SELECTING ESTABLISHMENTS WITHIN SAMPLE PSUs

Sample establishments within PSUs were selected independently in each size class using a systematic selection procedure. 5,983 establishments were selected in the initial sample. Establishments within sample PSUs were stratified by size class and 4-digit SIC. Sampling rates were applied to select establishments with less than 2,500 employees. Since the sampling rate for establishments with greater than 2,500 employees was so high, selection was done across all establishments nation-wide in these categories sequenced by zip code and 4-digit SIC. A sample of establishments not included in the survey because less than eight employees were listed for them on the DMI was interviewed, and it was found that a loss of about 5.5 percent in coverage of these small establishments existed in the NOES. This under-coverage in NOES might have existed because of growth in the number of employees in these establishments between the time the DMI was compiled and the time of the survey.

Workload control for the field interview phase of NOES was accomplished by enlarging the initial sample by 25 percent and dividing each PSU into 4 random subsamples. Each subsample was to be assigned in sequence. This was done to minimize the chance that an untimely termination of the survey would result in a non-representative sample. The enlarged sample was called the screening sample or telephone screening sample because it was comprised of establishments to be screened by telephone to determine if they were eligible or not for the survey, and if they would participate in it. If an establishment would not participate in the NOES, a suitable substitute was to be found from a sample of reserve establishments. This reserve sample was called the shadow sample.

A. The Number of Establishments

The proportion of establishments to be selected in a given size class was determined by the sampling rate $f = n_a/N_a$, where n_a was the number of establishments in size class a in the sample, and N_a was the total number of establishments in size class a.

In the NOES, sampling rates in each size class were determined from the formula:

$$f = \frac{n_a}{N_a} = \frac{S(\bar{Y}_a)C}{\left[\left(\sum_a N_a S(\bar{Y}_a)\sqrt{C_a}\right)\right]\sqrt{C_a}}$$

derived in Chapter IV (equation 5) and in Appendix D. For convenience in presentation f is expressed in terms of the sampling fraction or sampling interval $k = 1/f$. The sampling fraction is the reciprocal of the sampling rate. If all establishments in size class a were arranged in a list, the sampling fraction or interval would indicate the number of establishments passed over between selections.

Sampling fractions in each size class a were compared by the oversampling ratio $f_a = k_1/k_a$. In the NOES sampling fractions in each stratum were compared to the value k_1 for employee size class 1. The oversampling ratio is the ratio of sampling rates in each size class and indicates how much more frequent sample selection in a given size class is compared to another, in this case employee size class 1. Since fewer establishments were found to be sampled from in each successively larger size class, the sampling rate increased in each employee size class. K_a, the inverse of the sampling rate, decreased in each size class. F_a is the ratio of the sampling rate in size class a to the sampling rate in size class 1 (the size class with the lowest sampling rate) and increased with increasing sampling rates across size classes. Since f_a increased with employee size class, sample selection for establishments was proportional to size.

Values of cost C_a and variation $S(\bar{Y}_a)$ (from prior experience in the NOHS) were used to calculate sampling rates. Values of N_a, C_a, n_a, k_a, and f_a by employee size class calculated for the NOES are shown in Table 1 of Chapter IV. The sampling rates shown in Table 1 were calculated assuming that the CBP data most accurately reflected the national economy. DMI counts were not used in determining sampling rates because the DMI was known to contain listings for out-of-business firms, duplicate listings, etc.

It should be noted that according to the NOES design, selection of sample establishments with less than 2,500 employees was restricted to sample PSUs. The oversampling ratio f_a shown in Table 1 assumed selection from the total number of U.S. establishments in a given size class, however, and to obtain equal probability of selection among all establishments within each size class in the sample PSU, selection probabilities should also have taken into account the probability of selecting the sample PSU from its stratum. Selection probabilities for establishments in strata a are completely defined by the condition:

$$\frac{f_a}{k} = \left(\frac{M_{hj}}{M_h}\right) x \left(\frac{M_h}{M_{hj}} \times \frac{f_a}{k}\right)$$

equation 7 in Chapter V. The first term on the right side of the equation represents the probability of selecting the PSU from its stratum, and is 1.0 for establishments in self-representing strata, while the second term shows the probability of selecting establishments from the a^{th} size class within the PSU.

B. Selecting Establishments

1. General Plan

Sample establishments within employee size classes were selected in each of the 98 sample PSUs using a systematic selection procedure. Systematic selection was used in order to insure that sampling in each size class would be done proportional to the total number of establishments in each size category.

The 98 PSUs were first arranged in order and establishments within PSUs were stratified by employee size class. Establishments within size strata were then sequenced by 4-digit SIC code. Systematic selection in each size category was carried out using k_a (the sampling interval in size class a). The first establishment in each size category was chosen using a random number table. The next (k_a-1) establishments on the sequenced list were skipped, and the next establishment on the list was chosen as the next sample establishment. The process was repeated choosing every k_a^{th} establishment until the end of the list was reached. Note that, since the selection procedure was carried out across PSUs in a given size category, k_a was considered to be constant in each size category. This procedure was followed for size classes 1-8 and 11. For classes 9 and 10, systematic selection was applied to a list of all large U.S. establishments in each of the two size classes 2,500-4,999 and 5,000+ employees. The list for these two size classes was sequenced by zip code within 4-digit SIC.

2. Establishments in Size Classes 1-8 and 11

Systematic sampling of establishments across PSUs was done in size classes 1-8 and 11. Since the oversampling ratios f_a were considered to be constant over PSUs for each size class a, the complete set of selection probabilities in these size classes was defined when the PSU selection probabilities (M_{hj}/M_h), oversampling factors f_a, and sampling interval k_a were known. Values of f_a are shown in Table 1, values of k_a are shown in Table 1 and Table 3, and values of the PSU selection probabilities are shown in Appendix B. Size class 11 refers to those establishments for which the number of employees was not reported in the DMI, but which were reported as operating in a target SIC. Including these establishments in the survey posed a problem in defining sampling rates because the sampling rates for size classes 1-10 were derived using CBP counts and experience from the previous NOHS survey, neither of which gave any indication of the expected numbers or time required to survey firms whose number of employees were not available. It was decided to group these firms in a separate size class (size class 11) and sample them at rates equivalent to size class 1.

In self-representing (SR) PSUs, the PSU selection probabilities are 1.0, so that the selection probability of sample establishments within these PSUs was f_a/k_1. In non-self-representing (NSR) PSUs, the probability of selection for establishments within the PSU was:

$$\left(\frac{M_h}{M_{hj}}\right) \times \left(\frac{f_a}{k_1}\right)$$

3. Establishments in Size Classes 9 and 10

The proportion of establishments to be selected from size classes 9 and 10 was so large that sample efficiency would have been impaired if sampling were confined to the sample PSUs. For example, the probability of selection for establishments in size class 10 was, from Table 1 in Chapter IV:

$$\frac{f_a}{k_1} = \left(\frac{91.100}{199.53}\right) = \left(\frac{1}{2.190}\right)$$

Many of the sample PSUs in the NOES were selected with probabilities smaller than 1/2.190 (see Appendix B). For establishments in size class 10 to have been selected from the sample PSUs, however, at least 1/2.190 (about 46 percent) of the establishments in size class 10 should have been included in each sample PSU. If the sample of size class 10 establishments had been restricted to sample PSUs, it would not have been possible to obtain the desired sampling rates even if all size class 10 establishments within the PSU had been included in the survey. The problem was similar, although not as severe, for establishments in size class 9 (2,500-4,999 employees).

Location was therefore not considered in selecting the sample for the two largest size classes. Systematic selection was done across all U.S. establishments, sequenced by zip code and 4-digit SIC code. Many of these large establishments were located in or near a sample PSU, however, and could be surveyed by a team working in a nearby PSU.

4. Establishments with Fewer than Eight Employees

Although the NOES was limited to those establishments on DMI lists reporting eight or more employees and operating within the set of target SICs, rejecting facilities with seven or fewer employees as out of scope could have introduced bias into the survey. This could have occurred since the DMI employee reports were not current. Establishments not eligible to be included in the survey according to the DMI could have grown by the time of the survey to the point that they were eligible for inclusion in it. Under-coverage of smaller establishments could have been possible.

To measure this potential source of bias, a sample of 200 establishments reporting seven or fewer employees was screened over the telephone. This screening was done to determine the current number of employees and whether or not current activities were within the target SICs.

Results from the 200 telephone interviews are shown in Table 2. Eleven of these establishments actually had 8 or more employees and operated within the target SICs. This suggests a loss of about 5.5 percent in coverage of these small establishments in

TABLE 2. TELEPHONE INTERVIEWS OF 200 ESTABLISHMENTS REPORTING SEVEN OR FEWER EMPLOYEES ON THE 1980 DMI FILE NOES 1981-1983

Telephone interview reports	Number
Total	200
Non-working phone	50
No answer	3
Out of business	6
Less than 8 employees	123
Non-target SIC	2
Refusal	5
Reported 8 or more employees	11

the NOES. However, the estimation procedure used for national projections adjusted NOES levels to be consistent with the levels from the 1980 CBP (10); this reduced (although not entirely eliminated) the coverage bias.

C. Workload Control - Defining Shadow and Screening Samples

The sampling rates and expected sample sizes shown in Table 1 in Chapter IV are initial results calculated using assumptions discussed in Chapter IV, i.e., similar survey costs (as person hours) to NOHS, constant relvariance, and a field team of 21 surveyors for the survey period. These assumptions did not apply in every instance, however. For example, it was unrealistic to expect that survey teams would be equally proficient at all times, non-interview problems would not appear, the number of surveyors would remain constant over the survey period, or other scheduling problems would not arise. Some flexibility in the sample design was needed to account for problems arising during the course of the field work and to allow for the possibility that the surveyors might work faster than expected. An expanded 'screening' sample was selected and subdivided into a number of random subsamples for workload control, and a 'shadow' sample was selected in case of non-response. Screening and shadow samples were selected using DMI listings. The number of establishments selected in the initial screening and shadow samples, and their respective sampling intervals are shown in Table 3. Results if the sample could have been selected from CBP records are also shown for comparison.

The schedule for surveying sample establishments was based on a predicted length of stay determined from PSU person hour needs. All PSUs were to be covered during the expected two-year period for field work. To maintain such a schedule each team had to finish each of its survey assignments in the allotted time. However, the time per establishment was not identical in all PSUs. Since the period of time that could be spent in a PSU was fixed, a variable workload was necessary for an efficient field operation. This variable workload enabled supervisors to better react to problems in the field. The system was as follows:

1. The initial sample of 5,983 establishments was expanded by 25 percent to a total of 7,478 establishments. This sample was called the "screening sample" because it was comprised of establishments to be screened by telephone to determine if they were eligible for the NOES, and if they would participate.

2. Four random subsamples of the expanded sample were formed as follows:

   Subsample A = 1/2 of the expanded sample,
   Subsample B = 1/4 of the expanded sample,
   Subsample C = 1/8 of the expanded sample,
   Subsample D = 1/8 of the expanded sample,

3. At the beginning of its scheduled stay in the PSU each team was assigned a portion of the expanded workload to interview (e.g.,

TABLE 3. EXPECTED NUMBER OF ESTABLISHMENTS BY SIZE CLASS IN INITIAL, SCREENING, AND SHADOW SAMPLES NOES 1981-1983

Size class	Reported number of employees	From CBP[1]	Expected Number of Establishments From DMI Listings		
			Initial[2]	Screening	Shadow plus screening
1	8 - 19	1,190	1,393	1,742	3,483
2	20 - 49	914	1,073	1,341	2,681
3	50 - 99	675	785	981	1,961
4	100 - 249	838	1,003	1,253	2,507
5	250 - 499	512	604	755	1,510
6	500 - 999	344	409	511	1,023
7	1,000 - 1,499	123	163	204	407
8	1,500 - 2,499	108	142	177	355
9	2,500 - 4,999	94	124	155	309
10	5,000 and over	97	139	174	261
11	N/A	--	148	185	371
Total expected sample establishments		4,895	5,983	7,478	14,868
Sampling interval[3], k, in size class 1		199.53	199.53	159.62	79.81

1 Expected total sample at the U.S. level assuming it could be selected from a file of CBP establishment records for 1978. Also shown in Table 1 in Chapter IV.

2 Expected number of selections from the 1980 DMI file before eliminating duplications and out-of-scope cases.

3 For the Chicago PSU only, $k_1 = 212.13$, $k_2 = 169.7$ and $k_3 = 84.85$ for the initial, screening, and the screening plus shadow samples respectively.

subsamples A and C). The portions were chosen such that, over all PSUs, the total sample interviewed would approximate the number of establishments computed for the initial sample. The team was expected to survey all of the assigned subsamples during its stay in that PSU.

4. With the completion of the initial assignment in the PSU, the team supervisor was assigned additional subsamples where possible. All additional subsamples assigned had to be completed in the time originally fixed as the length of stay for the PSU.

Establishments included in the screening sample also had a reserve establishment selected with them for use in replacing attrition due to non-response. The sample of reserve establishments was called the shadow sample and was used as a substitute for non-cooperating establishments if all efforts during the telephone interview and by the surveyor and the team leader failed in obtaining cooperation from the establishment. The reserve was used as a substitute only for those original sample establishments currently in business and eligible for the survey. Furthermore, if the substituted shadow was found not to be eligible, or refused to cooperate, the initial sample unit was retained in sample and a court order (warrant) was obtained to secure cooperation from the originally designated unit. Original sample establishments found at the time of the survey to be out of business, or not doing business in any of the target SICs, were treated as ineligible and shadows were not substituted for them.

The values of f and k to determine sample size in the screening and shadow samples were computed by the methods discussed in Section A of this chapter and, except for classes 9 and 10, were based on a tabulation of the CBP establishment counts in the NOES target SICs for each size class. An early set of CBP counts (5) were used to derive the sampling rates by a clerical procedure before the more precise 1980 CBP counts (10) became available for size classes 9 and 10.

The screening sample was obtained by reducing the initial sampling interval, k_1, to $k_2 = (.8) * k_1$. For all size classes, except those reporting 5,000 or more employees, the screening sample and its reserve were designated in one operation by doubling the screening sample rate (that is, by using sampling intervals equal to half the intervals needed for the screening sample alone) and assigning alternate selections to the screening and shadow samples. The sampling intervals for the shadow and screening samples together for size classes 1-9, and 11 were then:

$$k_3 = (.5) * k_2 = (.5) * (.8) * k_1$$

Since the proportion of establishments to be selected from size class 10 was so high, the screening and shadow samples for it were selected with a systematic sampling interval 2/3 of the interval needed for the screening sample alone rather than 1/2 as for the other size classes. According to this, two sample establishments from size class 10 would share a single shadow establishment.

VII. THE FIELD INTERVIEW SAMPLE

Telephone screening of establishments in the screening sample was conducted to determine which of those establishments should be interviewed in the field. Telephone screening was intended to verify or correct basic information on sample establishments obtained from the DMI, collect further information, or modify the sample to include multi-facility establishments. 7,392 telephone interviews were conducted, and 4,850 establishments were found to be eligible for survey. Random subsamples A, B, C, D of the screened sample were determined for a variable field interview workload and were assigned individually to surveyor teams. The full A, B, C, D sample was interviewed in half of the selected PSUs, whereas interviews in subsamples A, B, and C were completed in the remaining PSUs. In all 4,490 establishments were interviewed in the field. The effective refusal rate for participation in the NOES was .3 percent.

A. The Field Interview Sample

The sampling scheme described in Chapters V and VI for selection of PSUs and selection of establishments within the PSUs provided the screening sample from which the field interview sample was derived. The screening sample was also referred to as the telephone screening sample since telephone screening was conducted on establishments in the screening sample to determine which establishments in this sample should be included in the survey, and should be interviewed in the field. The actual field interview operation was then accomplished most efficiently by dividing the workload into four random subsamples (see Chapter VI). The procedures followed during the field interview are discussed in Volume I of this series.

Field data for the NOES was collected after four steps:

1. A statistical sample of establishments was designated using the DMI file. The expanded sample (screening sample) and all shadows for each PSU were designated in one operation.

2. Telephone screening was carried out for the sample units. Telephone screening was intended to verify or correct basic information on sample establishments obtained from the DMI, and to collect further information. In addition, some screening information was used for sample modification. A single sample establishment might operate in more than one location or include several plants or branches, yet be listed only once on the DMI with a single address and employee total. If other establishments were owned or managed by the sample establishment, a search of the DMI file was done to determine if the new location should be treated as an addition to the sample. If the new location was not found on the DMI file, it was given a chance of selection to be included in the interview sample. In all, 93 multi-facility establishments were added to the screening sample in this way. Units not eligible for the survey which were identified during telephone screening were dropped from the survey.

3. Random subsamples A, B, C, D of the screened sample were determined. A variable interview workload was assigned to each surveyor team. The workload was designed so that each team could complete its assignment in a two week survey period. The assignment included subsamples A, B, C in size classes 1-8 and 11 and the full set of subsamples A, B, C, D in size classes 9 and 10. In roughly half of the PSUs it was possible to use the full sample by including subsample D. Expected times to complete the interviews are shown in Appendix B, and PSUs where the entire workload A, B, C, D was completed are shown in Appendices F (self-representing PSUs) and G (non-self-representing PSUs).

4. Field surveyors contacted each of the selected establishments to schedule an interview. A field surveyor visited each establishment, made a final determination of survey eligibility, and surveyed the establishment. Units determined not to be eligible at the time of the field survey were dropped from the study. If possible, substitutes from the shadow sample were found for eligible establishments refusing to participate in the study; if no substitute could be found, court warrants to require cooperation were obtained.

Results of the telephone screening interviews are shown in Table 4, and results from the field operations are shown in Table 5. Table 4 shows that 7,392 telephone interviews were conducted, of which 7,167 were of establishments included in the expanded screening sample and 225 were of establishments added because of refusals or determination of multi-facility establishments. Of the 7,392 establishments interviewed over the telephone, 4,850 (66%) establishments were found to be eligible for the field operations phase.

Each of the 4,850 establishments eligible for the survey were contacted for field interview. During the field interview, 346 establishments were found to be out-of-scope for the survey and 4,504 were determined to be in-scope (see Table 5). Only 4,379 (90%) of these in-scope establishments cooperated with a field interview, while 125 refused to cooperate. The shadow sample provided substitutes for 113 of these refusals, and warrants were used to complete the field operation in the remaining 12 establishments. Fourteen field interviews could not physically be completed during the survey period. This left 4,490 establishments for which field interviews for the NOES were completed.

The overall refusal rate for establishments to participate in either the telephone screening or field interview operations of the NOES was 7.1 percent. After substitution of establishments in the shadow sample for refusals and enforcing cooperation with court warrants, the effective refusal rate in the NOES dropped to .3 percent. The effective refusal rate was due to 14 establishments whose field interviews could not be completed during the survey period, and would better be described as the rate of non-response in the survey.

TABLE 4. RESULTS OF TELEPHONE SCREENING OPERATIONS
NOES 1981-1983

Telephone screening interviews	Screening sample	Added[1] sample	Total
Out-of-scope	2,535	7	2,542
Non-working phone	682	1	683
Out of business	230	2	232
Less than 8 employees	978	1	979
Non-target SIC	229	--	229
Govt. and administrative office	365	3	368
Out of PSU	51	--	51
In-scope	4,632	218	4,850
Refusals	221	--	221
Other In-scope	4,411	218	4,629
Total	7,167	225	7,392

1 Results from a subsample of 93 multi-facility establishments discovered during telephone screening.

TABLE 5. RESULTS OF FIELD OPERATIONS
NOES 1981-1983

Screening Field Operations	Added sample	sample[1]	Completed Total	interview	Not included[2]
Out-of-scope	339	7	346		
Out of business	64	2	66		
Less than 8 employees	186	1	187		
Non-target SIC	21	--	21		
Government	11	2	13		
Administrative office	39	1	40		
Work load subsamples	18	1	19		
In-scope	4,293	211	4,504	4,490	14
Cooperators	4,293	86	4,379	4,367	12
Subsampled plants	--	86	86	86	--
Screening sample establishments	4,293	--	4,293	4,281	12
Refusals	--	125	125	123	2
Shadows	---	113	113	111	2
Warrants	---	12	12	12	--
Total Field Operations	4,632	218	4,850	4,490	14

1 Results of a subsample of 93 multi-facility establishments discovered during telephone screening.

2 Could not be completed during survey period.

The overall attrition rate for establishments sampled, but found not to be eligible for inclusion in the study was 39.1 percent. This high value is due primarily to the expansion of the original sample by 25 percent for the telephone screening operation. This expanded sample for telephone screening was useful, however, to ensure that the sample of establishments actually surveyed in the field included enough eligible establishments to be as close as possible to the sample sizes calculated in Chapter VI. This feature of the sampling scheme minimized bias during the selection process. Non-response was so low as not to be a problem.

VIII. ESTIMATION PROCEDURES

National estimates of the number of employees and number of establishments conducting business in the SIC ranges covered by the NOES were obtained by assigning appropriate weighting factors to sample establishments and using these factors to project figures found in the NOES sample to the national level. A probability of selection was associated with each of the steps followed in determining the sample establishments which were interviewed. Inverses of these probabilities define sample weights which indicate how much each establishment's results contribute to national totals, and which can be used to provide estimates of the total number of establishments for the entire DMI file. Inflation estimates of totals were obtained by multiplying each establishment's totals by its sample weight and summing across establishments. These inflation estimates were followed by two stages of ratio estimation before the final publication estimate was determined. The first stage ratio estimation factor was based on establishment counts by employee size class as reported in the DMI. The second stage ratio estimation factor was based on employee counts (establishment counts for establishments with greater than 1,000 employees) by employee size class by SIC as reported in the CBP. Ratio estimation was used to improve the precision of the estimates.

Each estimate had a sample error associated with it. Furthermore, the complex survey design and estimation procedures used in the NOES lead to approximate and complicated expressions for estimation of the sampling error. Calculation of the sampling errors was handled using the method of replications. The method required that the estimation procedures be independently carried out several times (replicated) using subsamples of the original sample, and the variance of the replicate estimates be used to measure the variance of the full sample. Sampling error was taken as the square root of the variance. This system was flexible enough to provide measures of reliability for all tabulations planned for the NOES data.

National estimates of characteristics and of sampling errors were performed using computer software developed for this purpose.

A. Estimation of Totals

The inflation estimate was taken as an initial estimate of characteristics on the national level. Inflation weights, defined as the inverse of the probability of selecting the sample establishments from whom characteristics were to be estimated, were used to prepare unbiased estimates of characteristics for all sample establishments. If Y is a characteristic of all establishments, with y the value of that characteristic found in the sample, the simple inflation estimate Y_I would be:

$$Y_I = W \times y$$

where W is the inflation weight. For example, suppose y = 100 employees were reported working in a sample of establishments with probability of selection f = .05. The simple inflation estimate Y_1 of all employees working in the category from which the sample was selected would be $Y_1 = (1/.05) \times 100 = 2{,}000$ employees.

The simple inflation estimate involves only characteristics of the sample. If more information about the target population were available, more precise estimates for totals could be obtained. Ratio estimation uses independent sources of information about sample characteristics to determine an estimate which is often more precise than one determined from inflation estimates.

As an example, consider a characteristic x estimated from the sample, such as the number of employees in an industry surveyed in the NOES. Suppose X is a measure of the same characteristic but obtained from an independent source, such as the DMI. Then the ratio r = X/x may be used to alter the inflation estimate Y_1 described above. If Y_1 is an inflation estimate of a NOES item, the ratio estimate $Y_2 = Y_1 * r = W * y * r$ may be more precise than Y_1 alone. In ratio estimation the ratio (X/x) should vary in the same proportion as (Y/Y_1). It has been shown that, if values of X and Y are correlated, the estimate Y_2 will be more precise then Y_1 (10). In the NOES, two stages of ratio estimation were used, first using number of establishments and then using number of employees. The DMI and CBP listings of all establishments were used as outside sources in calculating the ratios.

1. Calculation of Inflation Weights

   In the NOES, inflation weights were determined in two stages. The telephone screening sample weight was first determined based on the sampling rates used to select the telephone screening sample, and then these weights were modified to take into account that portion of the sample actively used for field operations. Figure 2 shows the relationship between the telephone screening and field interview samples used to determine the weights, and Table 6 shows components of weights used in the NOES estimation procedure. Derivation of inflation weights (and ratio estimates) are also outlined there. The field interview weights were taken as the inflation weights used in the inflation estimates.

   A two step process was required in determining field interview weights since several of the survey operations had an impact on the exact values of the inflation weights and had to be accounted for. Telephone screening and the field interview operations both affected the sample weights, and it was simplest to correct for each phase separately. These operations involved:

   a. Assignment of variable workload subsamples to the PSUs.
   b. Sampling establishments with certainty in some PSUs.
   c. Results of the telephone interview.
   d. Substitution of shadow sample cases for refusals.
   e. Duplicate listing in the DMI file.

FIGURE 2. RELATIONSHIP BETWEEN TELEPHONE SCREENING AND FIELD INTERVIEW SAMPLES NOES 1981-1983

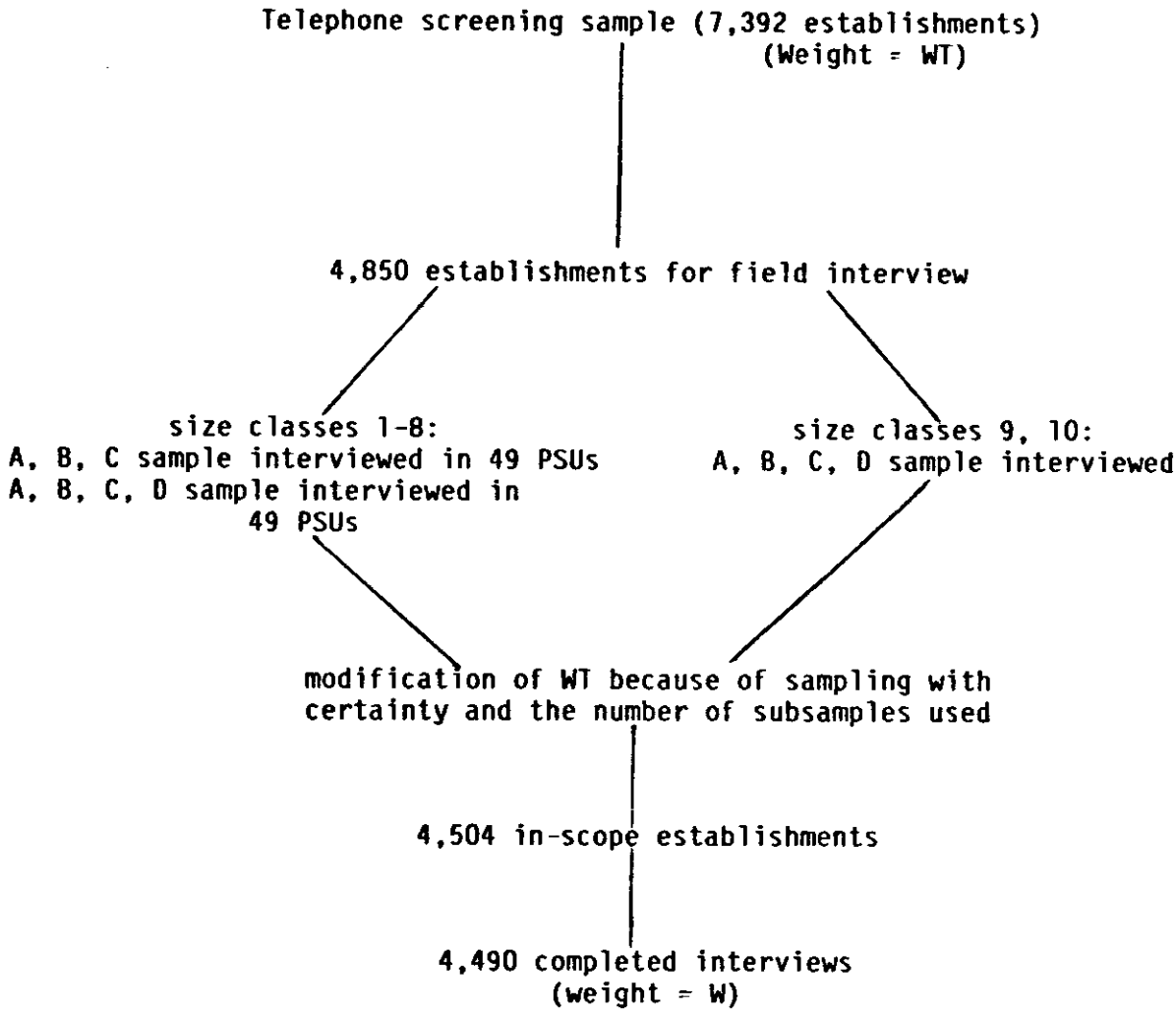


TABLE 6. COMPONENTS OF WEIGHTS USED IN THE NOES ESTIMATION PROCEDURE NOES 1981-1983

	Notation
Telephone screening sample weight:	WT
Assigned to each telephone sample establishment and based on the inverse of the probability of selecting DMI establishments in the sample.	
Field interview weight:	W
Assigned to each sample establishment interviewed in field, based on adjusted telephone sample weight.	
First stage ratio estimation factor:	R1
Numerator: DMI establishment counts by employee size class and area.	
Denominator: Estimates of numerator from telephone sample using weights = WT	
Second stage ratio estimation factor:	R2
Numerator: County Business Pattern employee counts (establishment counts for larger firms) by current size, and SIC.	
Denominator: Estimates of numerator from interviewed establishments using weights = W * R1	
First stage ratio estimates of characteristic from field interviews	W * R1
NOES estimates for publication	W * R1 * R2

Telephone sample weights were calculated considering a) and b) above, and the calculation of field interview weights considered points c), d), and e) above.

Subsamples A, B, C, D in each PSU were assigned for workload control as detailed in Chapter VI. Variable workloads consisted of either subsamples A, B, C or A, B, C, D of the expanded (screening) sample in each PSU, depending on which size class was being considered. Subsamples A, B, C were interviewed in size classes 1-8 and the full sample A, B, C, D was assigned in size classes 9 and 10. For half of the PSUs, however, it was possible to assign the full (sample) A, B, C, D for all size classes. Table 7 shows theoretical telephone sample weights by establishment size class for PSUs with A, B, C or A, B, C, D PSUs. Weights shown in Table 7 were calculated from counts of facilities appearing on the DMI, and so are different from the preliminary results shown in Table 1 in Chapter IV. Weights for the Chicago PSU are listed separately since sampling probabilities in that PSU were determined prior to a revision in sampling rates that occurred when more current CBP figures became available. PSUs assigned the full A, B, C, D sample in all size classes are listed in the Appendices G and H.

A second problem in determining telephone sample weights occurred because of sampling with certainty in some PSUs. This problem occurred with PSUs which could not meet the size criteria discussed in Chapter V. The probability of selection of establishments in these PSUs was so large in certain size classes that all establishments in those size classes would have been included in the sample: half of the establishments would be in the screening sample, and half in the shadow sample.

For example, consider samples selected from size class 5. From Table 7, the theoretical probability of selection of establishments from a PSU where the full A, B, C, D sample was interviewed would be 1 in 17.010 (the weights shown in Table 7 are inverses of the selection probabilities). Since a shadow sample equal in size to the full telephone sample was also selected in each PSU, the theoretical probability for establishments in size class 5 to be included in the combined screening and shadow sample would be: 2 x 1/17.010 = 1/8.005. The within PSU selection probability for establishments in size class 5 would then be: ((1/(probability of selection of sample PSU) x (1/1.8005)). This corresponds to the term $(M_h/M_{hj} \times f_a/k)$ in equation 7 in Chapter IV. Probabilities of selection for each of the 98 PSUs are shown in Appendix B.

The within PSU selection probability was less than 1 in most PSUs. For establishments in size class 5 in PSU 201, for example, the within PSU selection probability was 2.854 x (1/8.005) = .357. Consider another PSU, however, with a lower probability of selection. For PSU 206, the theoretical selection of probability of establishments in size class 5 would be: 11.984 x (1/8.005) = 1.50. If none of the assumptions made

TABLE 7. ESTABLISHMENT SIZE CLASSES AND THEORETICAL TELEPHONE SAMPLE WEIGHTS NOES 1981-1983

Size class	Number of employees	Theoretical weights for telephone sample		
		ABC PSUs	ABCD PSUs	Chicago
1	8-19	182.420	159.618	169.687
2	20-49	114.520	100.205	106.527
3	50-99	60.365	52.819	56.152
4	100-249	33.386	29.213	31.056
5	250-499	19.440	17.010	19.083
6	500-999	13.437	11.757	12.499
7	1000-1499	10.584	9.261	9.845
8	1500-2499	7.670	6.711	7.135
9[1]	2500-4999	---	4.436	---
10[1]	5000+	---	1.752	---
11	N/A	182.42	159.618	169.687

[1] ABCD samples selected from size classes 9 and 10 are not confined to sample PSUs.

in defining the sampling scheme had been violated, every establishment in size class 5 in PSU 206 should have been included as a sample establishment.

The problem of sampling with certainty in PSU 206 was solved in the NOES by taking half of the establishments in size class 5 in this PSU as the screening sample and half as the reserve sample, and assigning weights (2 x 11.984) = 23.968 for these establishments. The same problem would have occurred in size classes 6-8 in PSU 206 where the selection probabilities for establishments were even smaller, so half of the sample establishments in each of these size classes were included in the screening sample and half in the shadow sample, and weights of 23.968 were assigned to establishments in these size classes. Sampling establishments with certainty depended on the probability of selection of the PSU and size class. Other PSUs for which sampling with certainty was possible, and the weights determined for them, are shown in Appendix F.

Modifications to inflation weights in the telephone screening sample because of information collected during the screening operation were made on a case by case basis. The modified weights were the field interview weights. If the establishment was permanently out of business or ineligible for the NOES, a field inflation weight of 0 was assigned. If establishments were temporarily out of business, refused to interview, or were eligible for the NOES field interview inflation weights equal to their telephone sample weights were given. If establishments owned or managed other plant locations within the same PSU which were not originally included on the DMI list, these new locations were given a probability of selection. If sample establishments refused to participate in the NOES, establishments from the shadow sample were substituted for them. The weight of the refusal was set to 0, and the field interview weight for the shadow NOES was taken as the telephone screening sample weight.

The telephone screening operation might also have indicated duplicate listings of establishments in the DMI file. Searches for duplicate listings were confined to single PSUs. The field interview inflation weight for duplicates was modified depending on the size class of duplicate establishments: if size classes were the same for all of the 'n' duplicates the field interview weights were adjusted by 1/n; if the size class of the duplicate (non-sample) were less than that from the interviewed establishment, no adjustment was done; while if the size class of the (non-sample) duplicate was greater than that reported from the interview of the sample establishment, the field interview weight was taken as the inverse of the PSU selection probability. This last step was a compromise equivalent to treating the sample unit as though it had been selected with certainty within the PSU.

2. Ratio Estimation

Inflation estimates using the field interview inflation weights were modified with two stages of ratio estimation before using them to obtain national estimates. Ratios for each PSU size class were calculated using data from a subgroup of the 98 sample PSUs. These subgroups were defined to be as similar as possible within the PSU size class for which characteristics were to be defined. For the first stage ratio factor, PSUs were ordered and ratios (observed counts/estimated number) for establishments were calculated, adding results from each PSU one at a time. PSUs were added until the ratio fell between .3333 and 3.0000, and at least four plants contributed to the estimate of the denominator. The number of PSUs included in each ratio defined a ratio cell. The procedure was repeated starting with the next PSU, until ratio cells had been defined using all PSUs. If a ratio cell included results from the last PSU, yet the ratio factor did not meet the criteria above, it was combined with the previous ratio cell. For the second stage ratio factor, ratio cells were formed by including establishments in order by SIC code, and observed and estimated numbers of employees were used in the ratio.

For the first stage of ratio estimation, establishment size classes were as defined in Table 1 in Chapter IV. PSUs were ordered in pairs, determined so as to achieve homogeneity among size classes in adjacent PSUs in the listing. Homogeneity was considered in terms of proportion of employees in manufacturing, in large firms (1,000 or more employees), in the petroleum, chemical and rubber industries, and geography. The ordering was done in pairs for later use in the sampling error calculations. The ordering of NSR and SR PSUs are shown in Appendices G and H.

The first stage ratio numerator for size classes 1-8 and 11 was defined as the total number of establishments in PSUs in the ratio cell indicated on the DMI to be operating in the target SICs. The ratio denominator was taken as the inflation estimate of the numerator and was calculated from the telephone survey weights in the size class for sample PSUs included in the ratio cell. This ratio was denoted R1. The first stage ratio estimate of the characteristic would be W * R1.

A similar procedure was followed in obtaining first stage ratio estimates for size classes 9 and 10. The sample of establishments in size classes 9 and 10 were selected without regard to location with respect to a sample PSU, and so for each of these classes one area was defined for ratio estimation: all 50 states and the District of Columbia.

The second stage of ratio estimation used employee and establishment counts by size class and SIC from CBP information. It should be noted that CBP employee size classes are defined differently from those in the DMI; these are listed in Table 8. To make maximum use of the available data, then,

TABLE 8. EMPLOYEE SIZE CLASSES USED IN SECOND STAGE OF RATIO ESTIMATION (CBP SIZE CLASSES) NOES 1981-1983

Size Class	Number of Employees
1	8 - 9
2	10 - 19
3	20 - 49
4	50 - 99
5	100 - 249
6	250 - 499
7	500 - 999
8	1000 - 1499
9	1500 - 2499
10	2500 - 4999
11	5000+

three groups of establishments based on number of employees were considered in determining second stage ratio estimators: establishments reporting 10-999 employees and operating in a target SIC listed in group A in Appendix I, establishments reporting 1,000 or more employees and operating in a target SIC listed in group B in Appendix I, and all other NOES sample establishments. The iterative procedure used in determining ratio cells was similar to those used in determining first stage ratio cells, but characteristics of those establishments in ratio cells operating in particular SIC groups and employee size classes, rather than only in employee size classes, were combined.

Figure 3 outlines the second stage ratio estimation methods in the three groups of sample establishments. Numerators for Group A establishments were defined as the CBP count of employees in the establishments operating in the target SIC group(s) comprising the trial cell. The denominator was taken as the sample estimate of the number of employees in establishments operating in the target SIC group(s) included in the ratio cell. The denominator was determined using first stage tabulation weights as W * R1. The same conditions for the ratio as for the first stage ratio factors, i.e., ratio between .33 and 3.0, and denominator of the ratio being based on at least four establishments, were used. If the ratio cell was unacceptable, establishments in the adjacent SIC group (see Appendix I) were included in the ratio cell, and the procedure was repeated. If a final trial cell including the last SIC group was unsatisfactory, it was combined with the previous acceptable cell.

In some situations in Group A establishments, the number of employees in certain SIC groups could not be obtained from the CBP publication because of disclosure problems. Information about an individual establishment might have been revealed if the data were published. If disclosure were a problem, the average of ratio factors computed for other non-disclosure establishments in the same size class was assigned. SICs where disclosure was a problem are listed in Appendix I.

Procedures for Group B establishments were similar to those for Group A with the exception that numerators and denominators of ratios were based on observed and expected numbers of establishments rather than number of employees. The order of combination of SIC codes was the same as for those in Group A; however, the CBP count of these larger establishments could not be determined using a 2-digit SIC grouping and the procedures described below for Group C establishments were used. SIC codes where this was a problem are also indicated in Appendix I.

Ratios for Group C establishments were determined as average ratio factors of groups A and B. No iteration or testing to define suitable ratio cells was done. For sample establishments in size classes 2 through 11, the arithmetic average of ratio

FIGURE 3. SUMMARY OF SECOND STAGE RATIO ESTIMATION METHODS IN THREE GROUPS OF SAMPLE ESTABLISHMENTS NOES 1981-1983

<table>
<tr><td colspan="4">Number of employees at establishment when interviewed</td></tr>
<tr><td>8-
9</td><td>10-
19</td><td>20-
49 . . .</td><td>500-
999</td><td>1000-
1499</td><td>5000
+</td></tr>
<tr><td></td><td colspan="3">Group A Establishments

Ratio factors based on published counts and estimates of number of CBP employees.

Ratio cells determined by iteration.</td><td colspan="2">Group B Establishments

Ratio factors based on published counts on estimates of number of CBP establishments.

Ratio cells determined by iteration.</td></tr>
<tr><td colspan="6">Group C Establishments

Ratio factors based on averages of ratios computed for corresponding size classes in Groups A and B.

Iteration not involved.</td></tr>
</table>

factors determined for the corresponding size levels in Groups A and B was used. The average was taken as the sum of all ratios in the size class in Groups A and B divided by the number of ratios. For size class 1 the average ratio for establishments in size class 2 were used.

The second stage ratio estimation factor is denoted as R2. It was included in the record for each establishment. The final NOES publication weight included the field interview weight W, the first stage ratio estimation factor R1, and R2. The final publication weight was W * R1 * R2.

B. Estimation of Sampling Error

Sampling error refers to the deviation of actual values of the characteristic being studied from values of that characteristic estimated from the survey results. In the NOES, sampling errors were the result of estimates being based on results from a sample of the total number of plants. Statistically, sampling error is defined as the square root of the variance, where the variance for a group of independent observations $x_1 \ldots x_n$ with mean $\bar{x}$ is defined as: $\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)/(n-1)$. The sample variance for the NOES was defined in Chapter IV. Sampling error was taken as the square root of the variance.

The calculation of variances using standard statistical formulas available in computer software packages assumes that sample observations are collected independently of one another, and that a uniform sampling rate is used for the entire sample. In the NOES, however, this assumption does not apply. Since the sample selection scheme employed grouping by PSUs and stratification by size, characteristics of establishments in the same PSUs or size classes might have been correlated. These correlations could affect the reliability of projected statistics based on this sample. The variability in sampling rates among size strata and modifications imposed because of operational considerations discussed in Part A above also affected sampling errors.

The method of Balanced Repeated Replications (BRR) was used to estimate variances (9, 12, 13). This method is flexible enough to provide measures of reliability for all tabulations for the NOES data. The method was first used on a large scale in analyzing complex sample surveys of the Bureau of the Census. It is one of a class of methods of calculating variances by resampling from the sample many times. BRR in the NOES uses random subsamples generated by combinations of data from pairs defined by PSUs in each of the SR and NSR strata. Pairs are defined in the two largest size classes, size classes 9 and 10, irrespective of PSU. The same selection sampling strategies as used for the full sample were used in each of the random subsamples. Sixty-four pairs (half-samples) and 32 replicates were formed. The half-samples are the repeated replications of BRR. They are "balanced" because the replicates were determined from pairs.

It should be noted that difficulties in estimating sampling errors are not limited to the NOES. Similar problems occur in any large, complex survey, since practical and economical sampling schemes often use stratification and clusters of sample units as in the NOES. Furthermore, BRR is one of several statistical techniques which have been developed for estimating sample errors. BRR, however, is a basic method offering a great deal of flexibility through efficient use of independent replications, besides being used for calculating sampling errors. BRR is particularly useful with complicated statistics or for tests of significance.

The sample selection methods used for NOES result in variance estimates that are slightly biased (usually overestimates), regardless of the type of variance estimation used. In this sense estimation of the variance is a conservative estimate. These biases arise because only one PSU was selected from each of the 9 size strata considered, so some strata had to be combined in the variance estimation procedure to obtain meaningful results. Combining PSU's from different strata introduced an extra element of variance for the estimation procedure; two sample PSU's could have been selected from each stratum at, however, the cost of a decrease in efficiency of the sampling scheme. Also systematic sampling of establishments within each sample PSU was used to select establishments. Defining half-samples for BRR from samples already determined by a systematic selection procedure also tends to yield slight overestimates of the variance. These biases were not considered serious.

Half samples were first constructed by treating the PSUs as pairs. All of the original records were used, sequenced by identification number, size class, and selection probability, to reflect the original survey sampling process. An identification number was not used in sequencing size classes 9 and 10. Alternating selections of establishments within each of the 26 self-representing PSUs and size classes 9 and 10 defined 28 pairs. Pairs of non-self-representing PSUs were formed by alternating selection of PSUs (rather than of establishments). See Appendices H and I for details on how NSR PSUs were combined into pairs. Sixty-four half samples were therefore defined: 1 from each of the 26 SR PSU's (26), 36 from pairing of the NSR strata (36); and 1 each from size classes 9 and 10.

Replicates were defined as a 50 percent subsample of the total sample obtained by choosing one member from each of the paired half-samples. Individual establishments in each pair were given codes 0 or 1 to identify which establishment in the pair would be included in the replicate. In each SR PSU size class 1-8, the sum of the PSU identification number and size class number was found. If the sum was odd, odd numbered establishments were given code 1 and even-numbered establishments were given code 0. If the sum was even, even-numbered establishments were given 1 and odd-numbered establishments 0. Sample units in the PSU with code 1 comprised one member of the pair, while units with code 0 comprised the second member.

Pair members in size classes 9 and 10 were defined in a similar manner. Establishments in the telephone screening sample in each size group were sequenced by order of selection, and 1 was assigned to even-numbered records in each size class and 0 was assigned to odd-numbered records. Members in each NSR pair were defined when the NSR PSUs were paired. Sample establishments in each PSU were taken as the same members of the pair as was the PSU.

Thirty-two replicates for BRR were defined. Each replicate included one member from each of the 64 pairs. Which member to choose from each pair ('first' or 'second') was determined using a random number table (see Appendix J). Somewhat more precise estimates of the sampling error might have been obtained with greater numbers of replicates, however, 32 replicates were chosen for convenience and to reduce costs.

To use the technique, estimates of characteristics in each of the replicates were found using the ratio estimation procedures described for the full sample estimation process. Variances in each replicate were then found, and summed. Variances were found using the standard formula:

$$\text{Var}(x') = \left[\sum_{r=1}^{32} (x'_r - x'_0)^2\right]/r, \text{ where}$$

$\text{Var}(x')$ = variance of estimate of characteristic x

x'_r = estimate of the characteristic x made from the r^{th} replicate,

x'_0 = estimate of the characteristic x made from the full sample.

r = replicate number

Estimates x'_0 of characteristic x were calculated for each replicate using the same methodologies to calculate weights and ratio estimates as were used for the full sample. Since each replicate was a 50 percent subsample of the total sample, numerical values for the inflation weights used in the estimation procedures varied from those used for calculations with the entire sample. Half of the 32 weights (1 for each of the 32 replicates) were zero if the establishment was not in the replicate, and half were about twice as large as for the entire sample.

Variance, as defined above, is an absolute numerical measure of variation. Absolute measures, such as the standard deviation or standard error, have as units of measurement the units that the variable was expressed in. The magnitude of the variance is also a function of the magnitude of the characteristic of interest. Since establishments' characteristics could have varied so greatly depending on establishment size and SIC code, a relative measure of variation was needed for comparisons in the NOES. In the NOES, the relative measure of variation was the element relvariance,

$$V_{x'}^2 = \text{Var}(x')/(x')^2.$$

Relvariance was calculated across all 32 replicates:

$$V_{x'}^2 = \left(\frac{1}{32(x'_0)^2}\right)\left(\sum_{r=1}^{32} (x'_r - x'_0)^2\right)$$

Sampling error may be determined as:

$$\text{S.E. } (x') = \sqrt{\text{Var}(x')}$$

$$\text{S.E. } (x') = \sqrt{(V_{x'}^2)\ (x')^2}.$$

It should be noted that the methodology described thus far may be used to calculate total variances of estimated characteristics. The within PSU variance, the component of variance resulting because only selected establishments from each PSU were interviewed, is also of interest. Within PSU variance may be investigated using the same methods as for total variance, however, each element in each half sample should reflect the alternate selections for the 50 percent subsample. The modification comes in defining first or second elements in each NSR pair: first and second elements should be defined in each NSR PSU in the same manner as was done for SR PSUs, and the first and second elements of both PSUs in the NSR pair combined to determine members for the replicate. The difference between the total variance and the within PSU variance would be an approximation of the between PSU variance.

Calculation of estimates and sampling errors of estimates was done using specially developed software written in the FORTRAN computer programming language. Three files are input: a file of identifiers of establishments with user-specified questionnaire responses, a file of establishment weights, and a file of national estimates of total numbers of plants and employees covered in the NOES. The user-selected estimates may include totals, ratios, and other functions of data collected on the questionnaire. Output includes, by size category (number of employees) for each characteristic analyzed:

1. National estimate of number of plants and number of employees with the characteristic.

2. Standard error of each estimate.

3. Percentage of the total estimated number of plants or employee in the specified size category with the characteristic.

Table 9 is a portion of a tabulation of the output. This particular table shows final NOES estimates, and associated standard errors of number of plants and employees in plants with industrial hygiene services. The table also shows the percentage of the total number of plants or employees estimated in the NOES to have industrial hygiene services. Results are presented by employee size class and SIC code.

Standard errors may be used to construct confidence intervals about the estimates. For example, if the NOES could be conducted several times, roughly two-thirds of the resulting estimates of the total number of small-sized plants (8-99 employees) with industrial hygiene services would be within 3,923 of the 60,895 estimated or between 56,972 and 64,818. Similarly, 95% of the estimates would be within 7,846 (2 x 3,923) of 60,895. In other words, confidence intervals for estimate may be found as the sum or difference of the standard error and the estimate.

Final estimates from NOES data of numbers of facilities and numbers of employees will be included in forthcoming reports in this series.

TABLE 9. FINAL NOES ESTIMATES OF NUMBER OF PLANTS AND EMPLOYEES IN PLANTS WITH INDUSTRIAL HYGIENE SERVICES NOES 1981-1983

SIC CODE	PLANTS SMALL (8-99)	PLANTS MEDIUM (100-49)	PLANTS LARGE (≥500)	PLANTS TOTAL	EMPLOYEES SMALL (8-99)	EMPLOYEES MEDIUM (100-499)	EMPLOYEES LARGE (≥500)	EMPLOYEES TOTAL
07	1179*	70*		1249*	24945*	7009*		31954*
	(563)	(67)	...	(553)	(14547)	(6680)	...	(15011)
	21.2%	100.0%		22.2%	24.1%	100.0%		28.9%
13	991*	214*		1204*	14324*	35006*		49330*
	(558)	(171)	...	(562)	(7654)	(21376)	...	(21576)
	11.5%	21.0%		12.5%	6.9%	20.1%		11.9%
15	1429*	167*	20*	1616*	49434*	37900*	25647*	112981*
	(553)	(84)	(18)	(539)	(17794)	(23840)	(18373)	(31712)
	5.7%	15.1%	14.7%	6.2%	8.8%	19.1%	19.0%	12.6%
16	1424*	339*	17*	1780*	43652*	58573*	21017*	123241*
	(435)	(189)	(28)	(542)	(14416)	(38147)	(21480)	(49050)
	12.6%	30.3%	15.9%	14.2%	15.4%	28.2%	20.8%	20.8%
17	3428*	236*		3664*	80525*	58191*		138716*
	(981)	(144)	...	(1020)	(21288)	(33125)	...	(44002)
	5.9%	12.3%		6.1%	6.4%	17.6%		8.8%
20	3283*	1849*	272*	5404	122972	429923	274301*	777197
	(648)	(303)	(99)	(781)	(19568)	(74103)	(66896)	(78024)
	28.8%	57.5%	48.6%	35.6%	31.7%	63.8%	45.6%	50.0%
21	30*		10*	40*	1866		29132*	30998*
	(43)	...	(19)	(45)	(2685)	...	(29367)	(28820)
	100.0%		12.2%	36.6%	100.0%		26.0%	27.2%
22	186*	877	203*	1267	8092*	198415	186252*	392759
	(100)	(169)	(69)	(169)	(5323)	(41864)	(78950)	(73153)
	6.4%	55.2%	68.0%	26.3%	7.6%	57.2%	71.3%	55.0%
23	1501*	830	73*	2404	53894*	146667*	44124*	244685
	(557)	(190)	(47)	(558)	(14487)	(44058)	(30217)	(44375)
	12.0%	26.7%	30.3%	15.2%	12.7%	24.0%	21.2%	19.7%
ALL TARGET SICs	60895	21397	5318	87610	2055236	4425522	6635891	13116649
	(3923)	(1526)	(394)	(3989)	(113248)	(316227)	(416179)	(512752)
	13.6%	42.8%	56.3%	17.2%	18.5%	44.9%	53.2%	39.3%

* Standard error > 25% of the estimate. The estimate may be unreliable.
... No facilities observed.

REFERENCES

1. Occupational Safety and Health Act of 1970, Public Law 91-596.

2. U.S. Department of Health and Human Services. March 1988. National Occupational Exposure Survey: Volume I, Survey Manual DHHS (NIOSH) Publication No. 88-106.

3. Executive Office of The President - Office of Management and Budget 1972. Standard Industrial Classification Manual. Washington, DC. G.P.O. No. 041-00066-6.

4. U.S. Department of Health, Education, and Welfare. July 1977. National Occupational Hazard Survey (NOHS): Vol. II, Data Editing and Data Base Development. DHEW (NIOSH) Publication No. 77-213.

5. U.S. Bureau of the Census, 1980. County Business Patterns, 1978, United States: Appendix B. U.S. Department of Commerce G.P.O. No. 003-024-01662-1.

6. Dun's Marketing Index, Dun and Bradstreet, Inc., 1980.

7. Kish, L. 1965. Survey Sampling. John Wiley and Sons, New York.

8. Hanson, R., D. Ward, J. Edmonds, and J. Escatell, November 1980. National Occupational Exposure Survey (NOES). Final Report, Contract No. 210-80-0057. Westat, Inc., Rockville, Maryland.

9. Hansen, M.H., W.N. Hurwitz, and W.G. Madow. 1953. Sample Survey Methods and Theory: Volume II Theory. John Wiley and Sons, New York, p. 218.

10. U.S. Bureau of the Census, 1982. County Business Patterns, 1980, United States. U.S. Department of Commerce G.P.O. No. 003-024-05774-3.

11. Levy, P.S., and S. Lemeshow. 1980. Sampling for Health Professionals Lifetime Learning Publications, Belmont, California.

12. McCarthy, Philip J. April 1966. Replication: An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. PHS No. 1000 - Series 2 No. 14.

13. McCarthy, Philip J. January 1969. Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique. Vital and Health Statistics. PHS No. 1000 - Series 2 No. 31.

14. Kish, L. and M.R. Frankel. September 1970. Balanced Repeated Replications for Standard Errors. Journal of the American Statistical Association. (65), 331.

15. U.S. Department of Labor. Bureau of Labor Statistics. January 1981. Supplements to Employment, Hours, and Earnings, States and Areas. Data for 1981-1983. Employment and Earnings (Table B-2). G.P.O. No. 029-001-02822-9.

APPENDIX A. SIC CODES SURVEYED
NOES 1981-1983

Category	SIC Range
Agriculture	0700-0799
Oil and Gas Extraction	1300-1389
Construction, or Special Trade Contractor	1500-1799
Manufacturing	2000-3999
Transportation, Communications, Electric, Gas, or Sanitary Services	4000-4999
Wholesale Trade	5000-5199
Retail Trade	5200-5999
Specialized Services	7000-8999

APPENDIX B. 98 SAMPLE PSUs
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU State	Counties
Self-Representing PSUs				
142	5.49	1.0	NY	Nassau, Suffolk
371	8.22	1.0	WI	Milwaukee, Ozaukee, Washington, Waukesha
381	11.51	1.0	IN KY OH	Dearborn Boone, Campbell, Kenton Brown, Clermont, Hamilton, Warren
392	6.24	1.0	KY MO	Johnson, Wyandotte Cass, Clay, Jackson, Platte, Ray
511	3.79	1.0	MD VA DC	Clavert, Charles, Frederick, Montgomery, Prince George Arlington, Fairfax, Loudoun, Prince William, Cities of: Alexandria, Fairfax, Falls Church, Manassas, Manassas Park Washington
542	5.77	1.0	MD	Anne Arundel, Baltimore, Carroll, Harford, Howard, City of Baltimore
552	4.73	1.0	GA	Butts, Cherokee, Clayton, Cobb, DeKalb, Douglas, Fayette, Forsyth, Fulton, Gwinnett, Henry, Newton, Paulding, Rockdale, Spaulding, Walton
561	3.60	1.0	FL	Dade, Monroe
731	7.76	1.0	CA	Orange
742	3.19	1.0	CA	San Diego
752	5.64	1.0	CO	Adams, Arapahoe, Boulder, Denver, Douglas, Gilpin, Jefferson
761	4.09	1.0	WA	King, Snohomish

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU State	Counties
Self-Representing PSUs to be interviewed over two years				
110	10.85 13.94	1.0	NJ NY	Bergen Bronx, Kings, New York, Putnam, Queens, Richmond, Rockland, Westchester
120	4.66 5.27	1.0	NJ PA	Burlington, Camden, Gloucester Bucks, Chester, Delaware, Montgomery, Philadelphia
130	6.67 9.38	1.0	MA NH	Barnstable, Dukes, Essex, Middlesex, Nantucket, Norfolk, Plymouth, Suffolk Rockingham
150	4.92 2.95	1.0	NJ	Essex, Hunterdon, Morris, Somerset, Union
160	5.40 9.47	1.0	PA	Allegheny, Beaver, Washington, Westmoreland
310	14.77 13.66	1.0	IL	Cook, Dupage, Kane, Lake, McHenry, Will
320	10.51 15.81	1.0	MI	Lapeer, Livingston, Macomb, Oakland, St. Clair, Wayne
330	5.52 8.23	1.0	IL MO	Clinton, Madison, Monroe, St. Clair Franklin, Jefferson, St. Charles, St. Louis, City of St. Louis
340	3.34 2.11	1.0	MN WI	Anoka, Carver, Chisago, Dakota, Hennepin, Isanti, Ramsey, Scott, Washington, Wright St. Croix
350	6.94 5.46	1.0	OH	Cuyahoga, Geauga, Lake, Medina
520	6.52 5.04	1.0	TX	Collin, Dallas, Denton, Ellis, Hood, Johnson, Kaufman, Parker, Rockwall, Tarrant, Wise

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU State	Counties
530	3.41 4.02	1.0	TX	Brazoria, Chambers, Fort Bend, Harris, Libert, Montgomery, Waller
710	12.95 12.95	1.0	CA	Los Angeles
720	2.46 4.44	1.0	CA	Alameda, Contra Costa, Marin, San Francisco, San Mateo
Non-Self-Representing PSUs				
201	5.78	2.854	NY	Albany, Greene, Montgomery, Rennselaer, Saratoga, Schenectady
202	5.44	1.885	RI	Bristol, Kent, Newport, Providence, Washington
203	5.51	1.201	NY	Erie, Niagara
204	7.29	6.531	CT	New London, Windham
205	2.98	8.046	ME	Hancock, Nennebec, Knox, Lincoln, Waldo, Washington
206	4.74	11.984	PA	Blair
207	2.54	7.375	NY	Cattaraugus, Chautauqua
208	5.26	3.164	PA	Lancaster
209	7.13	1.973	CT	Fairfield
210	3.32	2.017	PA	Lackawanna, Luzerne, Monroe, Wyoming
211	2.49	2.882	NJ	Passaic, Sussex
212	3.85	5.954	NJ	Mercer
213	3.97	5.189	PA	Columbia, Montour, Schuylkill, Sullivan
214	3.66	2.227	NJ	Middlesex
401	7.37	8.879	MI	Genessee, Shiawassee

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU State	Counties
402	9.61	2.073	IN	Boone, Hamilton, Hancock, Hendricks, Johnson, Marion, Morgan, Shelby
403	2.53	2.872	IA NE	Pottawattamie Douglas, Sarpy
404	2.51	13.305	MN	Benton, Sherburne, Stearns
405	3.25	7.077	WI	Brown
406	2.30	16.050	KS	Douglas, Franklin, Leavenworth, Miami
407	2.34	8.835	OH	Guernsey, Harrison, Tuscarawas
408	5.05	1.787	OH	Delaware, Fairfield, Franklin, Madison, Pickaway
409	4.61	2.739	MI OH	Monroe Fulton, Lucas, Ottawa, Wood
410	5.84	3.762	IN	Adams, Allen, DeKalb, Wells, Whitley
411	2.49	13.362	MO	Audrain, Boone, Callaway, Howard, Randolph
412	1.74	16.814	KS MO	Allen, Anderson, Bourbon, Coffey, Linn, Woodson Barton, Bates, Henry, St. Clair, Vernon
413	3.89	8.535	WI	Racine
414	6.14	16.327	OH	Knox, Marion, Morrow
415	5.18	11.979	MI	Hillsdale, Lenawee
416	2.62	7.908	IN OH	Lagrange, Steuben Defiance, Henry, Paulding, Williams
417	5.22	16.355	IN	Dubois, Knox, Pike, Spencer
418	8.84	2.768	OH	Portage, Summit

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU	
			State	Counties
601	1.78	11.849	TX	Bee, Brooks, Dimmit, Duval, Frio, Goliad, Jim Hogg, Jim Wells, Karnes, Kennedy, Kinney, Kleberg, LaSalle, Live Oak, Maverick, McMullen, Starr, Uvalde, Willacy, Zapata, Zavala
602	1.99	2.507	FL	Broward
603	7.24	1.418	LA	Jefferson, Orleans, Plaquemines, St. Bernard, St. Charles, St. Tammany
604	2.26	16.870	TX	Atascosa, Bandera, Blanco, Bosque, Burnet, Caldwell, Comanche, Erath, Gonzales, Hamilton, Kerr, Medina, Mills, San Saba, Somervell, Wilson
605	2.11	13.643	TX	Austin, Bastrop, Colorado, Fayette, Jackson, Lavaca, Lee, Matagorda, Wharton
606	2.55	4.856	MS	Hinds, Madison, Rankin
607	1.67	9.920	TX	Clay, Montague, Wichita
608	2.75	1.196	FL	Hillsborough, Pasco, Pinellas
609	3.80	1.993	AR MS TN	Crittenden DeSoto Shelby, Tipton
610	7.96	2.052	OK	Creek, Mayes, Osage, Rogers, Tulsa, Wagoner
611	4.58	7.073	AL	Autauga, Elmore, Montgomery
612	4.70	4.703	SC	Lexington, Richland
613	2.57	3.564	AK	Pulaski, Saline
614	4.59	3.621	DE MD NJ	New Castle Cecil Salem

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU	
			State	Counties
615	4.99	17.158	VA	Dinwiddie, Prince George, Cities of: Colonial Heights, Hopewell, Petersburg
616	7.24	15.921	AL	Choctaw, Clarke, Conecuh, Monroe, Washington
617	3.95	20.721	SC	Clarendon, Georgetown, Williamsburg
618	4.03	12.059	NC	Johnson, Wilson
619	3.82	18.318	KY	Bath, Elliot, Fleming, Johnson, Laurence, Lewis, Magoffin, Martin, Mason, Menifee, Montgomery, Morgan, Nicholas, Robertson, Rowan, Wolfe
620	5.73	2.292	SC	Greenville, Pickens, Spartanburg
621	3.01	14.522	MD	Somerset, Wicomico, Worcester
622	5.33	1.920	NC	Davidson, Davie, Forsyth, Guilford, Randolph, Stokes, Yadkin
623	2.77	3.461	GA TN	Catoosa, Dade, Walker Hamilton, Marion, Sequatchie
624	4.39	9.234	AL	Calhoun, Etowah
625	4.52	21.775	VA	Bedford, Franklin, Rockbridge, Cities of: Bedford, Buena Vista, Lexington
626	3.73	10.201	OH WV	Washington Wirt, Wood
627	5.96	12.052	NC	Caswell, Granville, Person, Rockingham
628	5.50	21.284	MS	Clay, Lowndes, Webster

APPENDIX B. 98 SAMPLE PSUs (Cont.)
NOES 1981-1983

PSU number	Expected team-weeks*	PSU probability 1 in:	Composition of PSU	
			State	Counties
629	2.90	14.542	GA	Dawson, Fannin, Gilmer, Habersham, Lumpkin, Murray, Pickens, Rabun, Towns, Union
630	4.59	16.618	TN	DeKalb, Putnam, White
631	4.46	18.029	KY	Anderson, Bracken, Carroll, Franklin, Gallatin, Grant, Harrison, Henry, Owen, Pendleton, Shelby, Spencer, Trimble
801	2.29	2.969	CA	Placer, Sacramento, Yolo
802	1.83	7.163	CA	Kern
803	1.41	28.990	AK	Divisions of: Upper Yukon, Fairbanks, South East Fairbanks
804	2.68	5.363	NV	Clark
805	3.89	2.177	CA	Riverside, San Bernadino
806	2.39	4.933	CA	Fresno
807	5.17	1.871	OR	Clackamas, Multnomah, Washington, Yamhill
			WA	Clark
808	2.74	6.501	CO	El Paso, Pueblo, Teller
809	8.07	1.170	CA	Santa Clara

* Expected time to complete the sample of firms with less than 2,500 employees located in the sample PSU plus time to complete sample of larger firms located in or near the sample PSU.

APPENDIX C. COVERAGE OF DMI AND CBP FILES USED TO PROVIDE DETAILED INFORMATION ON SAMPLE ESTABLISHMENTS NOES 1981-1983

The adequacy of the DMI file was examined by comparing the total number of employees reported for target firms listed on DMI with corresponding totals from the CBP (5). Several problems occurred in comparing CBP and DMI tabulations:

1. The two files did not refer to the same time periods; CBP tabulations were for 1977 with establishment size classes in most cases based on the number of employees reported as of mid-March 1977. The DMI file was labeled "1980" with number of employees as carried on the most recent DMI record.

2. Establishments in scope for the study were confined to firms with eight or more employees. However, the CBP tabulations did not provide counts for the necessary establishment size classes so that approximations were required.

3. SIC coding for establishments was probably not consistent for the two files. For this reason, comparisons were made initially at the 2-digit SIC levels. Where serious differences appeared at the 2-digit level, the examination progressed to 3- and 4-digit levels. This assumed coding inconsistencies would be more evident at the detailed SIC levels.

4. CBP files exclude government employees, self-employed persons, farm workers, employees under the Railroad Retirement Act, and domestic service workers. About 24 percent of the total paid civilian wage and salary employment did not appear in the CBP tabulations. The absence of the self-employed was not considered a problem as they were assumed to be concentrated among firms too small to be in scope. The absence of the other categories may have accounted for some of the observed differences for the target SIC's.

The extent of coverage of government workers in DMI was not clear although a few government installations were found on the DMI universe lists. In some situations, the DMI file was evaluated using counts of employees on non-agricultural payrolls by industry as given in their Statements of Employment and Earnings (15); these figures referred to essentially the same group of employees as the CBP except that civilian government workers were included.

One criterion for the sampling design was that establishments from a file covering 90 percent or more of the target universe would be adequate for the study purposes. For establishment groups that did not meet this criterion, supplementing the DMI was considered. Supplementation would not be considered, however, unless the under-represented group of establishments comprised a workforce of at least 0.5 percent or so of the total 29,000,000 employees in all target establishments.

Comparisons of the DMI and CBP files indicated under-representation of the following SIC groups in the DMI (see also Chapter IV):

- 451 & 452 - Air transportation.
- 481 - Telephone communication.

- 491 - Electronic services.
- 493 - Combination electric, gas and other services combined.
- 5541 - Gasoline service stations.
- 7231 - Beauty shops.
- 7241 - Barber shops.
- 7299 - Miscellaneous personal services.

Supplementing the coverage of establishments in these SIC groups was considered using a second commercial list, the National Business List (NBL). However, the NBL could not provide the number of employees for each establishment and this information would have had to be obtained by telephone interview with each establishment selected.

In the case of gasoline service station attendants, for example, the NBL could have supplied a list of about 126,000 service stations that were not supposed to be on the DMI. The sample from this additional source would have been about 790 cases which would have had to be contacted by phone to screen out those with less than 8 employees; an expected 74 of these would have 8 or more employees and therefore be in scope (assuming all were still in business). The cost of adding the 74 additional cases to the sample would have been roughly $35 per case not counting the cost of telephone screening of the 790 units and the field interview cost of the 74 units. This sample supplement would also have to be matched against the DMI universe listing to remove any establishments already having a chance of selection, and selection probabilities for those added establishments would have to be found. Matching the NBL and DMI lists would also have had to be done before adding beauty shops, barber shops, or establishments performing miscellaneous personal services to the sample.

Since the NBL was constructed from essentially the same sources as the DMI, supplementing the remaining SIC groups (451, 452, 481, 491, 493) was not expected to be of much help in improved coverage. Further supplementation could also have been obtained by performing a search for firms appearing in phone directory yellow pages for the localities in the sample PSU's, but this project was considered beyond available resources. For these reasons, the coverage provided by the DMI was accepted without supplementation.

Oversampling establishments with employees in particularly hazardous occupations was also considered (e.g., construction). If a subset of establishments could have been identified as having higher rates of hazard exposures than other establishments, more reliable estimates for hazard exposures could have been obtained. If a subset of 10% of all establishments could have been identified as having exposure nine times as great as other remaining establishments, for example, it would be possible to reduce the sampling error for establishments exposed to that particular hazard by as much as 10 percent. This approach could not be adopted, however, because of problems in identifying high hazard exposure establishments, and the fact that oversampling for one characteristic might be a disadvantage when other characteristics were investigated.

APPENDIX D. DERIVATION OF SAMPLE SIZE FORMULAS NOES 1981-1983

Notation

The following notation is used:

Let N_a = the total number of establishments in the U.S. in all target industries in the a^{th} employee size class.

n_a = the number of establishments selected with equal probability from N_a.

C_a = the average cost (in person-hours) to investigate a sample establishment in the a^{th} employee-size category.

$C = \sum_a n_a C_a$ the cost of investigating the $n = \sum_a n_a$ sample establishments in terms of person-hours.

$\bar{y}'_a$ = the estimated average value of the y characteristic per establishment in the a^{th} size class based on the sample of n_a facilities in that class.

$y' = \sum_a N_a \bar{y}'_a$

$S^2(\bar{y}_a)$ = the estimated population variance of y'_a.

$= k(y'_a)^2$ (assumed)

$\sigma^2(y') = \sum_a N_a^2 (1/n_a - 1/N_a) S^2(\bar{y}_a)$

= the variance of the estimated total y'.

The optimum design for a sample may be determined using either the Cauchy Inequality or LaGrange Multipliers. Two basic quantities, C (total cost), and $\sigma^2(y)$ (variance of estimated characteristic) must be defined. Using the Cauchy Inequality, optimal sample size n_a in stratum 'a' with total number of numbers N_a at fixed cost C is found as the solution to the equation:

$$[\sigma^2(y')]^2 [\sqrt{C}]^2 = (\sigma^2(y'))(C)$$

or, substituting,

$$(\sum_a S_a^2(\bar{y}')/n_a)(\sum_a C_a n_a) = \sum_a (S_a^2(y')/n_a)(C)$$

The result is:

$$n_a = [N_a S(\bar{y}_a)/\sqrt{C_a}][(C_a)/\sqrt{C_a} N_a S(\bar{y}_a)]$$

LaGrange Multipliers may also be used. In this method, the variance function is constructed from the variance of the mean and variable cost determined by the LaGrange multiplier:

$$\phi = \sigma^2(y') + \lambda C$$

$$\phi = \sigma^2(y') + \lambda (a n_a c_a - C)$$

Partial derivatives of ϕ with respect to n_a are taken, the partial derivatives are set to ϕ, and the resulting simultaneous equations are solved for σ^2 and then for n_a.

For details on use of Cauchy's Inequality see Kish (14). See Hansen (9) for details on the LaGrange method.

APPENDIX E. DERIVATION OF FORMULA FOR A SELF-WEIGHTING SAMPLE NOES 1981-1983

To determine an expression defining a self-weighting sample first consider the overall probability of selecting a specific establishment. This probability is equal to (the probability of selecting the PSU containing the establishment) times (the probability of selecting the establishment from that PSU).

First define the following parameters:

M_{hj} = The total number of establishments for the survey in the j^{th} PSU and h^{th} stratum, i.e., $\sum_a N_{hja} f_a$, the measure of size of the j^{th} PSU in the h^{th} stratum.

M_h = The total number of establishments in the h^{th} stratum, i.e., $\sum_j M_{hj}$, the measure of size of all PSUs in the h^{th} stratum.

N_{hja} = The number of establishments in the U.S. in the a^{th} employee size class (according to CBP) in the hj^{th} PSU.

f_a = The oversampling ratio for establishments in the a^{th} size class (see below).

k = Sampling interval, $1/(n_a/N_a)$.

The probability of the PSU being selected in the J^{th} PSU and h^{th} stratum is M_{hj}/M_h. To obtain a self-weighting sample, establishments in the h, j^{th} PSU should be selected with a rate r_{hj} such that the sampling rate for establishments is proportional to the probability of the PSU being selected, or such that:

$$1/k = \left(M_{hj}/M_h\right) \times r_{hj}, \qquad (1)$$

where 1/k is the sampling fraction desired. From this,

$$r_{hj} = \left(M_h/M_{hj}\right) \times 1/k \qquad (2)$$

Substituting (2) into (1), a self-weighting sample may be defined by the condition:

$$1/k = \left(M_{hj}/M_h\right) \times \left(M_h/M_{hj} \times 1/k\right).$$

APPENDIX F. TELEPHONE SAMPLE WEIGHTS FOR ESTABLISHMENTS IN PSUs HAVING SIZE CLASSES SAMPLED WITH CERTAINTY NOES 1981-1983

PSU	SIZE CLASS					
	3	4	5	6	7	8
204				14.928	14.928	14.928
205				18.391	18.391	18.391
206			23.968	23.968	23.968	23.968
207				16.857	16.857	16.857
212				13.609	13.609	13.609
213					10.378	10.378
401			20.295	20.295	20.295	20.295
404			26.61	26.61	26.61	26.61
405				16.176	16.176	16.176
406		32.1	32.1	32.1	32.1	32.1
407			20.194	20.194	20.194	20.194
410						8.599
411			30.542	30.542	30.542	30.542
412		38.432	38.432	38.432	38.432	38.432
413			19.508	19.508	19.508	19.508
414		32.654	32.654	32.654	32.654	32.654
415			27.38	27.38	27.38	27.38
416				18.075	18.075	18.075
417		37.383	37.383	37.383	37.383	37.383
601			23.698	23.698	23.698	23.698
604		38.56	38.56	38.56	38.56	38.56
605			27.286	27.286	27.286	27.286
606					11.099	11.099
607			19.84	19.84	19.84	19.84
611				14.146	14.146	14.146
612					10.75	10.75
613						8.146
614						8.277
615		39.218	39.218	39.218	39.218	39.218
616		36.391	36.391	36.391	36.391	36.391
617		41.442	41.442	41.442	41.442	41.442
618			24.118	24.118	24.118	24.118
619		36.636	36.636	36.636	36.636	36.636
621			33.193	33.193	33.193	33.193
623						7.911
624			21.106	21.106	21.106	21.106
625		49.771	49.771	49.771	49.771	49.771
626			20.402	20.402	20.402	20.402
627			27.547	27.547	27.547	27.547
628		42.568	42.568	42.568	42.568	42.568
629			33.239	33.239	33.239	33.239
630		33.236	33.236	33.236	33.236	33.236
631		36.058	36.058	36.058	36.058	36.058
802				16.372	16.372	16.372
803	57.98	57.98	57.98	57.98	57.98	57.98
804					10.726	10.726
806					9.866	9.866
808				13.002	13.002	13.002

APPENDIX G. ORDER OF COMBINING SELF-REPRESENTING PSUs FOR FIRST STAGE RATIO ESTIMATION AND FOR VARIANCE ESTIMATION NOES 1981-1983

Pair number	PSU
1	110
2	150
3	120
4	142
5	130
6	160*
7	552*
8	542
9	381*
10	350
11	320*
12	371
13	310* #
14	520*
15	330
16	752*
17	392*
18	340*
19	720*
20	761*
21	530*
22	511
23	561*
24	742*
25	731*
26	710*
27	999* &
28	999* &

---

* Workload subsamples ABCD interviewed in the PSU (workload subsamples ABC in all other PSUs).

\# Within PSU selection probabilities differ from other PSUs, see text and Appendix D.

& Pairs 27 and 28 refer to size classes 9 and 10, respectively.

APPENDIX H. ORDER OF COMBINING NON-SELF-REPRESENTING PSUs FOR FIRST STAGE RATIO ESTIMATION AND FOR VARIANCE ESTIMATION NOES 1981-1983

Pair number	PSUs in pair: First member	PSUs in pair: Second member
29	601*	602*
30	801*	802
31	401	803*
32	804*	603*
33	604	605*
34	606	607*
35	805*	806*
36	402*	403
37	404*	608*
38	609	610
39	611*	807*
40	808*	405
41	201	202
42	612	613
43	406*	614
44	615	616
45	203	204
46	617*	618*
47	205	206*
48	619*	620
49	207	208*
50	209	210*
51	407	408*
52	409	410
53	411	621
54	622*	623
55	624	809*
56	211	625
57	412	212
58	413	626*
59	414*	213*
60	627	628*
61	415	416
62	629	630*
63	417	631*
64	214	418*

* Workload subsamples ABCD assigned in the PSU (workload subsamples ABC in all other PSUs).

APPENDIX I. ORDER FOR COMBINING 2-DIGIT SIC SUMMARIES TO SECOND STAGE OF RATIO ESTIMATION NOES 1981-1983

GROUP A: Establishments reporting 10-999 employees in the following SICs:

Order	SIC	Order	SIC
1	15[D]	21	36
2	16	22	37
3	17	23	38
4	20	24	39
5	21	25	41 (411, 412, 415, 417)
6	22	26	44
7	23	27	45
8	24	28	46[D]
9	25	29	48
10	26	30	49
11	27	31	50[D,L] (501, 503, 505, 5093)
12	28	32	51[D,L] (516, 517)
13	29	33	55 (552, 553, 554)
14	13	34	72[D] (not including 7218)
15	30	35	73[D,L] (733, 734, 7391, 7395, 7397, 7399)
16	31	36	75[D] (not including 752)
17	32	37	76
18	33	38	84
19	34		
20	35		

GROUP B: Establishments reporting 1,000 or more employees in SICs listed in group A.

Establishments reporting 1,000 or more employees in SICs 50, 51, 73 were assigned to Group C.

D CBP employee counts for one or more size classes will show "Disclosure".

L CBP count of large establishments (more than 1,000 employees) cannot be determined for the 2-digit SIC; Group C ratio procedure used for large firms.

APPENDIX I. ORDER FOR COMBINING 2-DIGIT SIC SUMMARIES TO SECOND STAGE OF RATIO ESTIMATION (CONT.) NOES 1981-1983

GROUP C: Establishments reporting 8 or 9 employees in any of the SICs enumerated in Groups A and B plus all establishments reporting 8 or more employees in the following SICs:

0723	422
0724	423
0742	4742
0782	478
0783	8062
4013	807
4212	809
4214	

Establishments reporting 1,000 or more employees in SICs 50, 51, 73.

APPENDIX J. RANDOM NUMBER TABLE USED TO DEFINE REPLICATES FOR VARIANCE ESTIMATION* NOES 1981-1983

REPLICATE	PAIR NUMBER 1				2				3				4				5				6	
	1234	5678	9012	3456	7890	1234	5678	9012	3456	7890	1234	5678	9012	3456	7890	1234	5678	9012	3456	7890	1234	
1	0000	1010	1110	1100	0111	1100	1101	0010	0000	1010	1110	1100	0111	1100	1101	0010						
2	1000	0101	0111	0110	0011	1110	0110	1000	1000	0101	0111	0110	0011	1110	0110	1000						
3	0100	0010	1011	1011	0001	1111	0011	0100	0100	0010	1011	1011	0001	1111	0011	0100						
4	0010	0001	0101	1101	1000	1111	1001	1010	0010	0001	0101	1101	1000	1111	1001	1010						
5	1001	0000	1010	1110	1100	0111	1100	1100	1001	0000	1010	1110	1100	0111	1100	1100						
6	0100	1000	0101	0111	0110	0011	1110	0110	0100	1000	0101	0111	0110	0011	1110	0110						
7	1010	0100	0010	1011	1011	0001	1111	0010	1010	0100	0010	1011	1011	0001	1111	0010						
8	1101	0010	0001	0101	1101	1000	1111	1000	1101	0010	0001	0101	1101	1000	1111	1000						
9	0110	1001	0000	1010	1110	1100	0111	1100	0110	1001	0000	1010	1110	1100	0111	1100						
10	0011	0100	1000	0101	0111	0110	0011	1110	0011	0100	1000	0101	0111	0110	0011	1110						
11	1001	1010	0100	0010	1011	1011	0001	1110	1001	1010	0100	0010	1011	1011	0001	1110						
12	1100	1101	0010	0001	0101	1101	1000	1110	1100	1101	0010	0001	0101	1101	1000	1110						
13	1110	0110	1001	0000	1010	1110	1100	0110	1110	0110	1001	0000	1010	1110	1100	0110						
14	1111	0011	0100	1000	0101	0111	0110	0010	1111	0011	0100	1000	0101	0111	0110	0010						
15	1111	1001	1101	0100	0010	1011	1011	0000	1111	1001	1101	0100	0010	1011	1011	0000						
16	0111	1100	1110	1010	0001	0101	1101	1000	0111	1100	1110	1010	0001	0101	1101	1000						
17	0011	1110	0111	0101	0000	1010	1110	1100	0011	1110	0111	0101	0000	1010	1110	1100						
18	0001	1111	0011	1010	1000	0101	0111	0110	0001	1111	0011	1010	1000	0101	0111	0110						
19	1000	1111	1001	1101	0100	0010	1011	1010	1000	1111	1001	1101	0100	0010	1011	1010						
20	1100	0111	1100	1110	1010	0001	0101	1100	1100	0111	1100	1110	1010	0001	0101	1100						
21	0110	0011	1110	0111	0101	0000	1010	1110	0110	0011	1110	0111	0101	0000	1010	1110						
22	1011	0001	1111	0011	1010	1000	0101	0110	1011	0001	1111	0011	1010	1000	0101	0110						
23	1101	1000	1111	1001	1101	0100	0010	1010	1101	1000	1111	1001	1101	0100	0010	1010						
24	1110	1100	0111	1100	1110	1010	0001	0100	1110	1100	0111	1100	1110	1010	0001	0100						
25	0111	0110	0011	1110	0111	0101	0000	1010	0111	0110	0011	1110	0111	0101	0000	1010						
26	1011	1011	0001	1111	0011	1010	1000	0100	1011	1011	0001	1111	0011	1010	1000	0100						
27	0101	1101	1000	1111	1001	1101	0100	0010	0101	1101	1000	1111	1001	1101	0100	0010						
28	1010	1110	1100	0111	1100	1110	1010	0000	1010	1110	1100	0111	1100	1110	1010	0000						
29	0101	0111	0110	0011	1110	0111	0101	0000	0101	0111	0110	0011	1110	0111	0101	0000						
30	0010	1011	1011	0001	1111	0011	1010	1000	0010	1011	1011	0001	1111	0011	1010	1000						
31	0001	0101	1101	1000	1111	1001	1101	0100	0001	0101	1101	1000	1111	1001	1101	0100						
32	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000						

* Entries are 0 or 1, depending on whether to include the second or first member, respectively, of each pair in the replicate.

★U.S. GOVERNMENT PRINTING OFFICE: 1990-749-701/H5696